

Optimal algorithms for universal random number generation from finite memory sources

Gadiel Seroussi, *Fellow, IEEE*, and Marcelo J. Weinberger *Fellow, IEEE*

Abstract—We study random number generators (RNGs), both in the fixed to variable-length (FVR) and the variable to fixed-length (VFR) regimes, in a universal setting in which the input is a finite memory source of arbitrary order and unknown parameters, with arbitrary input and output (finite) alphabet sizes. Applying the method of types, we characterize essentially unique optimal universal RNGs that maximize the expected output (respectively, minimize the expected input) length in the FVR (respectively, VFR) case. For the FVR case, the RNG studied is a generalization of Elias’s scheme, while in the VFR case the general scheme is new. We precisely characterize, up to an additive constant, the corresponding expected lengths, which include second-order terms similar to those encountered in universal data compression and universal simulation. Furthermore, in the FVR case, we consider also a “twice-universal” setting, in which the Markov order k of the input source is also unknown.

I. INTRODUCTION

Procedures for transforming non-uniform random sources into uniform (“perfectly random”) ones have been a subject of great interest in statistics, information theory, and computer science for decades, going back to at least [1]. For the purposes of this paper, a *random number generator* (RNG) is a deterministic procedure that takes, as input, samples from a random process over a finite alphabet \mathcal{A} , and generates, as output, an integer r that is uniformly distributed in some range $0 \leq r < M$ (where M may also be a random variable, depending on the setting, in which case uniformity is conditioned on M). If $M = p^m$, the output r can be regarded as the outcome of m independent tosses of a *fair p -sided coin* (or *die*); when $p = 2$, it is often said, loosely, that the RNG generates m *random bits*.

Various assumptions on the nature of the input process, what is known about it, and how the samples are accessed, give raise to different settings for the problem. Regarding the sample access regime, we are interested in two such settings. In the *fixed to variable-length* RNG (in short, FVR) setting, it is assumed that the input consists of a fixed-length prefix X^n of a sample of the random process, and the size M of the output space depends on X^n . The *efficiency* of the FVR scheme is measured by the expectation, with respect to the input process, of $\log M$,¹ the *output length*, which we seek to *maximize*. In other words, the goal is to obtain as much output “randomness” as possible for the fixed length of input consumed. In the *variable to fixed-length* RNG (in

short, VFR) setting, the size M of the output space is fixed, but the length n of the input sample is a random variable. The set of input sequences for which such a scheme produces an output is referred to as a *stopping set* or *dictionary*. In this case, efficiency is measured by the expectation of n , the *input length*, which we seek to *minimize*. The goal is to consume as few samples of the input process as possible to produce a pre-determined amount of “randomness.”

An RNG is said to be *universal* in a class of processes \mathcal{P} if, conditioned on the size M of the output range, it produces a uniformly distributed output for any process in the class (conditioning on M is trivial in the VFR case). We are interested in universal RNGs that optimize efficiency (i.e., maximize expected output length in the FVR case, or minimize expected input length in the VFR case), simultaneously for all processes $P \in \mathcal{P}$. FVRs have been studied extensively, in various universal (e.g. [2], [3]) and non-universal (e.g., [4]) settings. VFRs were studied in [1],² [5], [6] (with emphasis on the case $M = 2$ and Bernoulli sources), and more recently, in more generality, in [7], [8], [9].

Although, in principle, the input length of a VFR is unbounded, we are also interested in *truncated VFRs* (TVFRs). A TVFR either produces a uniformly distributed output on an input of length $n \leq N$, for some fixed N , or it *fails* (producing no output). We require VFRs to produce uniform outputs, while admitting some failure probability, *at all truncation levels* N . In that sense, our notion of a universal VFR is stricter than the one in the earlier literature (cf. [1], [5], [6], [7], [8]), where generally no conditions are posed on the truncated VFRs. The stricter notion may be useful in practical applications, where there is likely to be some prior knowledge or minimal requirements on the statistics of the source (e.g., some assurance of minimal “randomness”). If the VFR has still not produced an output after consuming an length of input for which this prior assumption implies that (with high probability) it should have, the whole system function may be considered suspect (for example, somebody might have maliciously impaired the random source). With the stricter definition, the input length threshold can be set arbitrarily, while preserving perfect uniformity of the VFR output. Although this may seem too fine-grained a restriction, it will turn out that the penalty it imposes on the expected input length of the optimal VFR is negligible relative to the

Gadiel Seroussi is with Universidad de la República, Montevideo, Uruguay (e-mail: gseroussi@ieee.org).

Marcelo J. Weinberger is with Center for the Science of Information, Stanford University, Stanford, CA 94305 (e-mail: marcwein@ieee.org).

¹Logarithms are to base 2, unless specified otherwise.

²The scheme in [1] can also be interpreted as an FVR, since it outputs at most one random bit per pair of input symbols. This is the point of view taken, e.g., in [3]. In general, VFRs can be used to construct FVRs, and vice versa. However, the resulting constructed RNGs will be generally less efficient than if they were optimized for the intended regime from the start.

main asymptotic term.

It is well known that in both settings of interest, the entropy rate \bar{H} of the source determines fundamental limits on the asymptotic performance of the RNGs, namely, an upper bound on the expected output length in the FVR case ($n\bar{H}$; see, e.g., [4]), and a lower bound on the expected input length in the VFR case ($(\log M)/\bar{H}$; see, e.g., [7] for i.i.d. sources). Furthermore, for various cases of interest, these bounds have been shown to be asymptotically tight to their main term. We are interested in a more precise characterization of the performance of optimal algorithms, including the rate of convergence to the fundamental limits. As in other problems in information theory, this rate of convergence will depend on the class of processes \mathcal{P} in which universality is considered. In this paper, we focus on the class \mathcal{P}_k of k -th order *finite memory (Markov)* processes, where the order k is arbitrary.

In the FVR case, we extend the notion of universality, and say that an FVR is *twice-universal* in the class $\mathcal{P} = \bigcup_k \mathcal{P}_k$ of *all* finite memory processes, with both k and the process parameters unknown, if its output approaches a uniform distribution for every process in the class (the formal definition of distribution proximity is provided in Section III). We seek twice-universal FVRs that attain essentially the same efficiency as universal FVRs designed with knowledge of the order k . The relaxation of the uniformity requirement is necessary to satisfy this strict efficiency goal, as will follow from the characterization of universal FVRs in Section III. To keep the paper to a reasonable length, we omit the study of twice-universality in the VFR case, in which the study of the basic universal setting is already rather intricate.

The contributions of the paper can be summarized as follows. In Section III, we first review results on universal FVRs. In [2], Elias presented a universal FVR procedure for Bernoulli processes and binary outputs (i.e., in our notation, the class \mathcal{P}_0 with $|\mathcal{A}| = 2$ and M a power of 2). The procedure is optimal in expected output length, pointwise for every input block size n and every Bernoulli process. An efficient implementation is described in [10], and a generalization to first-order processes for any \mathcal{A} is presented in [11].³ Although not explicitly employing the terminology, these schemes can be seen as implicitly relying on basic properties of *type classes* [12], [13]. We show that Elias’s procedure, when studied explicitly in the more general context of the *method of types* [12], [13], is applicable, almost verbatim and with a uniform treatment, to broader classes of processes, and, in particular, to the class \mathcal{P}_k for any value of k and any finite alphabet \mathcal{A} , while retaining its universality and pointwise optimality properties. We precisely characterize, up to an additive constant, the expected output length of the procedure in the more general setting. The estimate shows that, similarly to “model cost” terms in universal compression and universal simulation schemes, FVRs exhibit a second-order term (a term

keeping the expected length below the entropy of the input) of the form $\frac{K}{2} \log n + O(1)$, where $K = |\mathcal{A}|^k(|\mathcal{A}| - 1)$ is the number of free statistical parameters in the model class \mathcal{P}_k . However, somewhat surprisingly, we observe that this term is incurred for almost all processes P in the class, even if the FVR is designed to produce uniform outputs only for P . Thus, in the case of FVRs, the second-order term is not necessarily interpreted as a “cost of universality,” as is the case in the other mentioned problems. After reviewing universal FVRs, in Subsection III-D we present a twice-universal FVR, also inspired on Elias’s scheme, but based on the partition, presented in [14], of the space \mathcal{A}^n into classes that generalize the notion of type class for the twice-universal setting. We show that the expected output length of the twice-universal FVR is the same, up to an additive constant, as that of a universal FVR designed for a known order k , with the (unnormalized) distance of the output to a uniform distribution vanishing exponentially fast with the input length.

We then shift our attention, in Section IV, to VFRs that are universal in \mathcal{P}_k , for arbitrary k . After formally defining the setting and objectives, in Subsection IV-C we characterize the essentially unique optimal VFR, which minimizes both the expected input length and the failure probability, pointwise for all values of M and at all truncation levels N (and, thus, also asymptotically). This VFR appears to be, in its most general setting, new, and it can be regarded as an analogue to Elias’s scheme, but for the variable to fixed-length regime. We precisely characterize, up to an additive constant, the expected input length of this optimal VFR, and show that, as its FVR counterpart, it also exhibits a second order term akin to a “model cost.” In the VFR case, the term is proportional to $\log \log M$ (the logarithm of the output length), and again to the number of free statistical parameters of the input model class. We also show that the failure probability of the optimal VFR decays exponentially fast with the truncation level N . In addition, we show that the scheme admits an efficient sequential implementation, which is presented in Subsection IV-E. We note that the optimal VFR coincides with the optimal “even” procedure of [5] for Bernoulli processes and $M = 2$. A universal VFR that, although sub-optimal for all N , asymptotically attains the entropy limit for Bernoulli processes and $M = 2^m$ was previously described in [9] (without an analysis of the second order term of interest here). The dictionary of [9] is a special case of an auxiliary construct we use, in Subsection IV-G, in the derivation of an upper bound on the expected length of the optimal VFR. The optimal scheme itself, however, is not derived from this construct.

In the computer science literature, FVRs are referred to as *randomness extractors* (RE), and a deep and extensive theory has developed, with connections to other fundamental areas of computation (see, e.g., [15] and references therein). We note that the approach we take to the problem (which follows traditional information-theoretic and statistical methodology) and the approach taken in the RE literature are rather different. The results obtained in the RE literature are very powerful in the sense that they make very few assumptions on the nature of the imperfect random source, other than a guarantee on the minimum information content of an input sample, namely,

³The class studied in [11] includes the class of k -th order Markov processes, if we interpret a process of order k over \mathcal{A} as a process of order 1 over \mathcal{A}^k . However, the knowledge that only $|\mathcal{A}|^{k+1}$ state transition probabilities are nonzero (out of $|\mathcal{A}|^{2k}$ in the generic case of this interpretation) can be used with obvious complexity advantages in the management of transition counts. In addition, this knowledge is necessary for the precise asymptotics derived in this paper.

a lower bound on $\min_{x^n} \{-\log P(x^n)\}$, referred to in the literature as the *min-entropy* of the source. Results on expected output length are then expressed in terms of this bound, and are usually tight up to multiplicative constants. By focusing on specific classes of input sources, on the other hand, we are able to obtain tighter results (with performance characterized up to additive constants), and, from a practical point of view, schemes where statistical assumptions can be reasonably tested (e.g., entropies can be estimated), which does not appear to be the case with the very broad assumptions employed in RE theory.

II. DEFINITIONS AND PRELIMINARIES

A. Basic objects

Let \mathcal{A} be a finite alphabet of size $\alpha = |\mathcal{A}|$. We denote a sequence (or string) $x_i x_{i+1} \dots x_j$ over \mathcal{A} by x_i^j , with x_1^j also denoted x^j , and $x_i^j = \lambda$, the empty string, whenever $i > j$. As is customary, we denote by \mathcal{A}^n and \mathcal{A}^* , respectively, the set of strings of length n , and the set of all finite strings over \mathcal{A} , and we let $x^* \in \mathcal{A}^*$ denote a generic string of unspecified length. For an integer $M > 0$, we let $[M] = \{0, 1, \dots, M-1\}$, and for integers t and u , we write $t \equiv u \pmod{M}$ if $M|(u-t)$, and $t = u \bmod M$ if $t \equiv u \pmod{M}$ and $0 \leq t < M$. We denote by \mathbb{N} and \mathbb{N}^+ the sets of nonnegative and positive integers, respectively.

An α -ary *dictionary* \mathcal{D} is a (possibly infinite) prefix-free subset of \mathcal{A}^* . We say that \mathcal{D} is *full* if and only if every string $x^n \in \mathcal{A}^n$ either has a prefix in \mathcal{D} , or is a prefix of a string in \mathcal{D} . Naturally associated with a dictionary \mathcal{D} is a rooted α -ary *tree* $\mathbf{T}_{\mathcal{D}}$, whose nodes are in one-to-one correspondence with prefixes of strings in \mathcal{D} . The leaves of $\mathbf{T}_{\mathcal{D}}$ correspond to the elements of \mathcal{D} , and, for $u \in \mathcal{A}^*$ and $a \in \mathcal{A}$ such that ua is a node of $\mathbf{T}_{\mathcal{D}}$, $\mathbf{T}_{\mathcal{D}}$ has a branch (u, ua) labeled with the symbol a . We identify nodes with their associated strings, and say that u is the *parent* of ua , or, conversely, that ua is a *child* of u . Moreover, we sometimes regard a tree as a set of sequences and say, e.g., that \mathcal{D} is the set of leaves of $\mathbf{T}_{\mathcal{D}}$. We note that the definition of a full dictionary is consistent with the usual definition of a *full tree* (sometimes also referred to as a *complete tree*). Clearly, \mathcal{D} is full if and only if every node of $\mathbf{T}_{\mathcal{D}}$ is either a leaf, or the parent of α children.⁴ In the sequel, all dictionaries (and corresponding trees) are assumed to be full. Notice that a full infinite tree need not satisfy the Kraft inequality with equality [16]. We will observe in Section IV, however, that if $\mathbf{T}_{\mathcal{D}}$ does not have a Kraft sum of one, then \mathcal{D} is rather useless for the construction of an efficient VFR. The set of internal nodes of $\mathbf{T}_{\mathcal{D}}$ is denoted $\mathbf{I}_{\mathcal{D}}$. A finite tree is *balanced* if it has α^ℓ leaves at its deepest level ℓ .

B. Finite memory processes and type classes

A k -th order *finite memory* (Markov) process P over the alphabet \mathcal{A} is defined by a set of α^k conditional probability mass functions $p(\cdot|s) : \mathcal{A} \rightarrow [0, 1]$, $s \in \mathcal{A}^k$, where $p(a|s)$ denotes the probability of the process emitting a immediately

after having emitted the k -tuple s . The latter is referred to as a *state* of the process, and we assume for simplicity a fixed initial state s_0 .

Let $b \in \mathcal{A}$ be a fixed arbitrary symbol. We denote by \mathbf{p} the parameter vector $\mathbf{p} = [p(a|s)]_{a \in \mathcal{A} \setminus \{b\}, s \in \mathcal{A}^k}$, and by Ψ its domain of definition. To simplify the statement of some arguments and results, we further assume that all conditional probabilities $p(a|s)$ (including $p(b|s)$) are nonzero, i.e., we take Ψ as an open set by excluding its boundary.⁵ The dimension of \mathbf{p} is equal to the number of free statistical parameters in the specification of the k -th order process P , namely, $K \triangleq (\alpha - 1)\alpha^k$.

The probability assigned by P to a sequence $x^n = x_1 x_2 \dots x_n$ over \mathcal{A} is

$$P(x^n) = \prod_{t=1}^n p(x_t | x_{t-k}^{t-1}), \quad (1)$$

where we assume that $x_{-k+1}^0 = s_0$. In cases where we need to consider a different initial state s , we denote the probability $P(x^n|s)$. The entropy rate of $P \in \mathcal{P}_k$ is denoted $\bar{H}_P(X)$ and is given by

$$\bar{H}_P(X) = \sum_{s \in \mathcal{A}^k} P^{\text{st}}(s) H_P(X|s) \quad (2)$$

where, for a state s , $P^{\text{st}}(s)$ denotes its stationary probability (which, by our assumptions on \mathcal{P}_k , is well defined and nonzero), and $H_P(X|s)$ denotes its conditional entropy, given by $-\sum_{a \in \mathcal{A}} p(a|s) \log p(a|s)$.

The *type class* of x^n with respect to the family \mathcal{P}_k of all k -th order finite memory processes is defined as the set

$$T(x^n) = \{y^n \in \mathcal{A}^n \mid P(x^n) = P(y^n) \ \forall P \in \mathcal{P}_k\}. \quad (3)$$

Let $n_s^{(a)}(x^n)$ denote the number of occurrences of a following s in x^n , i.e.,

$$n_s^{(a)}(x^n) = |\{t \mid x_{t-k}^t = sa, 1 \leq t \leq n\}|, \quad a \in \mathcal{A}, s \in \mathcal{A}^k,$$

and $n_s(x^n) \triangleq \sum_{b \in \mathcal{A}} n_s^{(b)}(x^n)$. Denote by $\mathbf{n}(x^n)$ the vector of α^{k+1} integers $n_s^{(a)}(x^n)$ ordered according to some fixed convention. It has been well established that the definition (3) is equivalent to the combinatorial characterization

$$T(x^n) = \{y^n \in \mathcal{A}^n \mid \mathbf{n}(x^n) = \mathbf{n}(y^n)\}.$$

The vector $\mathbf{n}(x^n)$ is referred to as the *type* of x^n .

The set of all type classes for sequences of length n is denoted \mathcal{T}_n , i.e.,

$$\mathcal{T}_n = \{T(x^n) \mid x^n \in \mathcal{A}^n\}.$$

The following fact about type classes $T \in \mathcal{T}_n$ is well known.

Fact 1: All sequences in T share the same final state, i.e., for some fixed string $u^k \in \mathcal{A}^k$, we have $x_{n-k+1}^n = u^k$ for all $x^n \in T$.

⁵In fact, our results only require that Ψ be a positive volume subset of the $\alpha^k(\alpha-1)$ -dimensional simplex. The additional requirement of excluding the boundary guarantees the validity of our asymptotic expansions for every parameter in Ψ .

Unless explicitly stated otherwise, we will assume that the RNG constructions described in this paper have access to the (arbitrary) initial state s_0 , and the order k of the processes, which are necessary to constructively define the type class partitions \mathcal{T}_n . We will depart from these assumptions in Subsection III-D when we discuss twice-universal RNGs (and the order k is not assumed known), and in Subsections III-C and IV-C when we briefly discuss RNGs that are insensitive to the initial state, and are based on a slightly different definition of the type class.

Type classes of finite memory processes (and of broader model families) have been studied extensively (see, e.g., [12], [13] and references therein). In particular, the cardinality of a type class is explicitly characterized by *Whittle's formula* [17], and a one-to-one correspondence between the sequences in a type class and Eulerian cycles in a certain digraph constructed from $\mathbf{n}(x^n)$ was uncovered in [18]. Whittle's formula also allows for the computationally efficient enumeration of the type class, i.e., the computation of the index of a given sequence in its class, and the derivation of a sequence from its index, by means of enumeration methods such as those described in [19].⁶ These enumerations are a key component of the RNG procedures discussed in this paper.

III. UNIVERSAL FIXED TO VARIABLE-LENGTH RNGS

A. Formal definition

An FVR is formally defined by a triplet $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ where $n \in \mathbb{N}^+$ is the fixed input length, $\mathbb{N}_t \subseteq \mathbb{N}^+$ is a *target set* such that $1 \in \mathbb{N}_t$, and $\rho : \mathcal{A}^n \rightarrow \mathbb{N}$, $\mathcal{M} : \mathcal{A}^n \rightarrow \mathbb{N}_t$, are functions such that $\rho(x^n) \in [\mathcal{M}(x^n)]$. The *output length* of \mathcal{F}_n on input x^n is defined as $\log \mathcal{M}(x^n)$. Thus, the function \mathcal{M} determines the range of the output random number and the output length, while the function ρ determines the random number itself within the determined range. When the goal is to generate fair p -sided coin tosses, we choose

$$\mathbb{N}_t = \{p^i \mid i \geq 0\}, \quad p \geq 2. \quad (4)$$

An FVR \mathcal{F}_n is *perfect* for a process $P \in \mathcal{P}_k$ if $\rho(x^n)$, conditioned on $\mathcal{M}(x^n) = M$, is uniformly distributed in $[M]$; \mathcal{F}_n is *universal* in \mathcal{P}_k if it is perfect for all $P \in \mathcal{P}_k$. The *expected output length* of $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ with respect to P is

$$L_P(\mathcal{M}) \triangleq \mathbf{E}_P \log \mathcal{M}(X^n) = \sum_{x^n \in \mathcal{A}^n} P(x^n) \log \mathcal{M}(x^n), \quad (5)$$

where \mathbf{E}_P denotes expectation with respect to P . Given a process order k , the goal is to find universal FVRs that maximize L_P simultaneously for all $P \in \mathcal{P}_k$. We are interested in L_P in a pointwise sense (i.e., for each value of n), and also in its asymptotic behavior as $n \rightarrow \infty$.

Notice that our setting is slightly more general than the usual one for FVRs in the literature, where the condition (4) for some p is generally assumed in advance. As we shall

see, there is not much practical gain in this generalization. However, the broader setting will allow us to better highlight the essence of the optimal solutions, as well as connections to related problems in information theory such as universal compression and universal simulation.

For conciseness, in the sequel, except when we discuss twice-universality in Subsection III-D, when we say “universal” we mean “universal in \mathcal{P}_k for a given order k , understood from the context.”

B. Necessary and sufficient condition for universality of FVRs

The following condition for universality is similar to, albeit stronger than, conditions previously derived for problems in universal simulation [22], [23], [14] and universal FVRs [24], [11]. The proof is deferred to Appendix A.

Lemma 1: Let $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ be an FVR satisfying the following condition: For all $T \in \mathcal{T}_n$ and every $M \in \mathbb{N}_t$, the number of sequences $x^n \in T$ such that $\mathcal{M}(x^n) = M$ and $\rho(x^n) = r$ is the same for all $r \in [M]$ (in particular, the number of sequences $x^n \in T$ such that $\mathcal{M}(x^n) = M$ is a multiple of M). Then, \mathcal{F}_n is universal in \mathcal{P}_k . If \mathcal{F}_n does not satisfy the condition, then it can only be perfect for processes P with parameter \mathbf{p} in a fixed subset Ψ_0 of measure zero in Ψ .

The following corollary is an immediate consequence of Lemma 1. It shows that universality is essentially equivalent to perfection for a single, “generic” process in \mathcal{P}_k .⁷

Corollary 1: An FVR is universal if and only if it is perfect for any single process $P \in \Psi \setminus \Psi_0$, where Ψ_0 is a fixed subset of measure zero in Ψ .

C. Variations on the Elias procedure

We start by considering the simplest target set, namely $\mathbb{N}_t = \mathbb{N}^+$ (i.e., no restrictions such as (4) are placed on the ranges of the generated random numbers). Let $\mathcal{I}_T(x^n)$ denote the index of $x^n \in \mathcal{A}^n$ in an enumeration of $T = T(x^n)$. The following procedure defines an FVR $\mathcal{F}_n^* = (\mathbb{N}^+, \rho^*, \mathcal{M}^*)$.

Procedure E1: Given an input sequence x^n , let $\mathcal{M}^*(x^n) = |T(x^n)|$, and $\rho^*(x^n) = \mathcal{I}_T(x^n)$.

Procedure E1 is a “bare-bones” version of Elias's procedure.⁸ It is straightforward to verify that \mathcal{F}_n^* satisfies the condition of Lemma 1 and is, thus, universal in \mathcal{P}_k . The following theorem shows that \mathcal{F}_n^* attains the maximum possible expected output length of any universal FVR for the given n , all $P \in \mathcal{P}_k$, and arbitrary target set \mathbb{N}_t .

Theorem 1: If $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ is universal in \mathcal{P}_k , then, for any $P \in \mathcal{P}_k$,

$$L_P(\mathcal{M}) \leq L_P(\mathcal{M}^*) = \mathbf{E}_P \log |T(X^n)|. \quad (6)$$

Proof: The equality is straightforward from the definition of Procedure E1. The inequality follows from Lemma 1 and Corollary 1, which imply that $\mathcal{M}(x^n) \leq |T(x^n)|$ for all $x^n \in \mathcal{A}^n$. ■

⁷In particular, this result settles a conjecture put forth in [25, p. 917].

⁸An equivalent procedure is described in [25] as a first step in the implementation of Elias's procedure, the second step consisting of a “binarization” of $\rho^*(x^n)$, for the case $p = 2$ in (4).

⁶In this context, “computationally efficient” means computable in polynomial time. Although further complexity optimizations are outside the scope of this paper, various tools developed for similar enumeration and RNG problems in the literature would be applicable also here, and should allow for significant speed-ups. See, e.g., [20], [21], [10], [11].

The term on the rightmost side of (6) was precisely estimated in [22] in the context of universal simulation of sources in \mathcal{P}_k , by analyzing the expectation of Whittle's formula, and obtaining

$$E_P \log |T(X^n)| = H_P(X^n) - (K/2) \log n + O(1), \quad (7)$$

where $H_P(X^n)$ denotes the entropy of the marginal $P(X^n)$, $K = (\alpha - 1)\alpha^k$, and the $O(\cdot)$ notation refers to asymptotics in n .

Remark 1: The second-order term $(K/2) \log n$ on the right hand side of (7) resembles a typical “model cost” term in universal lossless compression. By Theorem 1, this term determines the rate at which the expected output length of \mathcal{F}_n^* approaches (from below) $H_P(X^n)$, which is the best possible convergence rate for *any* universal FVR. Notice, however, that by Corollary 1, the bound of Theorem 1 applies even if the FVR is required to be perfect just for a single process with parameter $\mathbf{p} \in \Psi \setminus \Psi_0$. Therefore, in this case, the second-order term must be incurred (almost always) also in the non-universal (known P) setting, and, in fact, it can be argued that there is *no asymptotic cost* for universality. Nevertheless, we will still refer to the second order term as a *model cost*, since it is proportional to the size of the model, regardless of whether the parameters of the input process are known or not.

Remark 2: Procedure E1 is similar to a *universal enumerative encoder*, a two-part universal lossless compressor for the class \mathcal{P}_k . The encoder differs from the FVR in that it outputs, together with $\rho^*(x^n)$ and in lieu of M , an efficient description of $T(x^n)$. It is known (see, e.g., [26]) that $K \log n + O(1)$ bits are sufficient for this description, resulting in an overall expected code length of $H_P(X^n) + (K/2) \log n + O(1)$, which is optimal, up to an additive constant, for any universal lossless compressor for the class \mathcal{P}_k . The rate of convergence to the entropy is the same as for FVRs, but convergence, in this case, is from above. We observe that, *a fortiori*, a universal lossless compressor *cannot be* a universal FVR for \mathcal{P}_k (and vice versa).

We now shift our attention to more general target sets, which include also sets of the form (4). Let \mathbb{N}_t be an arbitrary subset of the positive integers with $1 \in \mathbb{N}_t$. For any $M \in \mathbb{N}^+$, let

$$\lfloor M \rfloor_{\mathbb{N}_t} = \max \{ j \in \mathbb{N}_t \mid j \leq M \}.$$

Let c be a constant, $c \geq 1$. We say that \mathbb{N}_t is *c-dense* if and only if for every $M \in \mathbb{N}^+$, we have

$$M \leq c \lfloor M \rfloor_{\mathbb{N}_t}.$$

For example, \mathbb{N}^+ is 1-dense (no other subset of \mathbb{N}^+ is), and the target set in (4) (used in Elias's procedure for fair p -sided coins) is p -dense. In the sequel, we will assume that \mathbb{N}_t is c -dense for some c .

Procedure E2 in Fig. 1 defines an FVR $\mathcal{F}_n^{**} = (\mathbb{N}_t, \rho^{**}, \mathcal{M}^{**})$ (we recall that $\mathcal{I}_T(x^n)$ denotes the index of x^n in an enumeration of $T(x^n)$). The assumption that $1 \in \mathbb{N}_t$, and the fact that $r < \mu$ holds throughout after the execution of Step 1, guarantee that Procedure E2 always stops, and, by the stopping condition in Step 2b, the output satisfies the required condition $r \in [M]$. It is also readily verified that \mathcal{F}_n^{**} satisfies the condition of Lemma 1 and is, thus, universal.

Input: Sequence $x^n \in \mathcal{A}^n$.
Output: Pair (r, M) , $M \in \mathbb{N}_t$, $r \in [M]$.

- 1) Let $\mu = |T(x^n)|$, $r = \mathcal{I}_T(x^n)$.
 - 2) Repeat forever:
 - a) Let $M = \lfloor \mu \rfloor_{\mathbb{N}_t}$.
 - b) If $r < M$ then output (r, M) and **stop**.
 - c) Let $\mu = \mu - M$, $r = r - M$.
-

Fig. 1. Procedure E2: Generalized Elias procedure (\mathcal{F}_n^{**}).

We refer to Procedure E2 as “greedy,” since, in Step 2c, it always chooses to reduce μ by the largest possible element of \mathbb{N}_t . The procedure trivially coincides with Procedure E1 when $\mathbb{N}_t = \mathbb{N}^+$. When \mathbb{N}_t is of the form (4), the procedure coincides with Elias's original scheme in [2], suitably extended to finite-memory sources of arbitrary order, and coins with an arbitrary number of sides.

Suppose we do not let Procedure E2 stop in Step 2b, instead allowing it to run until $\mu = 0$ in Step 2c. Then, the procedure defines a decomposition of $|T(x^n)|$ into a sum

$$|T(x^n)| = \sum_{i=1}^m M_i, \quad (8)$$

where $M_i \in \mathbb{N}_t$, and $M_1 \geq M_2 \geq \dots \geq M_m$. The term M_i corresponds to the value that M assumes at the i -th execution of Step 2a, namely,

$$M_i = \left\lfloor |T(x^n)| - \sum_{j=1}^{i-1} M_j \right\rfloor_{\mathbb{N}_t}, \quad 1 \leq i \leq m, \quad (9)$$

where m is the first index such that $M_m \in \mathbb{N}_t$ (m is well defined since $1 \in \mathbb{N}_t$).

Remark 3: Equations (8)–(9) define a partition of the integer $|T(x^n)|$ into summands in \mathbb{N}_t , which, through an enumeration of $T(x^n)$, translates to a partition of $T(x^n)$ into subclasses, with the size of each subclass belonging to \mathbb{N}_t . This partition induces a refinement of the original type-class partition of \mathcal{A}^n , so that all the sequences in a refined subclass are still equiprobable for all $P \in \mathcal{P}_k$. Procedure E2 can then be interpreted as applying Procedure E1, but using the refined partition in lieu of the type-class partition. The procedure stops when it finds the subclass x^n is in, at which time the value of r is the index of x^n in the subclass.

Next, we characterize the expected output length of \mathcal{F}_n^{**} when \mathbb{N}_t is c -dense. The characterization will make use of the following technical lemma, a proof of which is deferred to Appendix B.

Lemma 2: Let $\mathbf{q} = [q_1, q_2, \dots, q_m]$, with $q_1 \geq q_2 \geq \dots \geq q_m > 0$, be the vector of probabilities of a discrete distribution on m symbols, and let $H = -\sum_{i=1}^m q_i \log q_i$ denote its entropy. Assume that for some constant $c \geq 1$, \mathbf{q} satisfies

$$c q_i \geq 1 - \sum_{j=1}^{i-1} q_j, \quad 1 \leq i \leq m. \quad (10)$$

Then, letting $h(\cdot)$ denote the binary entropy function, we have $H \leq c h(c^{-1})$.

Theorem 2: If \mathbb{N}_t is c -dense, the expected output length of \mathcal{F}_n^{**} for $P \in \mathcal{P}_k$ is

$$L_P(\mathcal{M}^{**}) = L_P(\mathcal{M}^*) - O(1). \quad (11)$$

Proof: Let $T = T(x^n)$ denote an arbitrary type class and let M_1, M_2, \dots, M_m denote the integers in \mathbb{N}_t determined by the decomposition of $|T|$ in (8)–(9). Define

$$\mathbf{q}(T) = |T|^{-1} [M_1, M_2, \dots, M_m]. \quad (12)$$

Clearly, $\mathbf{q}(T)$ is the vector of probabilities of a discrete distribution, with entropy $H(\mathbf{q}(T))$. By (9), the c -density assumption applied to the quantities $|T| - \sum_{j=1}^{i-1} M_j$, $1 \leq i \leq m$, and the definition (12), Lemma 2 applies to $\mathbf{q}(T)$. Since the sequences in a type class are equiprobable, the expectation of $\log \mathcal{M}^{**}(X^n)$ conditioned on T is given by

$$L(T) \triangleq |T|^{-1} \sum_{i=1}^m M_i \log M_i = \log |T| - H(\mathbf{q}(T)), \quad (13)$$

and

$$\begin{aligned} L_P(\mathcal{M}^{**}) &= \sum_{T \in \mathcal{T}_n} P(T) L(T) \geq E_P \log |T| - c h(c^{-1}) \\ &= L_P(\mathcal{M}^*) - O(1), \end{aligned} \quad (14)$$

where the inequality follows from (13) and Lemma 2, and the last equality follows from the rightmost equality in (6). The claim of the theorem now follows by combining the lower bound (14) with the upper bound in Theorem 1. ■

Remark 4: In a worst-case sense, the sufficient condition of c -density in the theorem is also necessary, since, using the fact that $L(T) \leq \log M_1$ by (13) (with the notation in the proof), we have

$$\log |T| - L(T) \geq \log |T| - \log M_1 = \log \frac{|T|}{|T|_{\mathbb{N}_t}}. \quad (15)$$

Thus, if \mathbb{N}_t is not c -dense for any c then the expression on the rightmost side of (15) is unbounded.

Theorem 2 shows that if \mathbb{N}_t is c -dense, \mathcal{F}_n^{**} performs to within an additive constant of the expected output length of \mathcal{F}_n^* , which is an upper bound for any universal FVR, independently of the target set. In particular, this implies that \mathcal{F}_n^{**} is optimal, up to an additive constant, among all FVRs for the same target set \mathbb{N}_t . While, for a general c -dense target set, this additive constant is positive, the following theorem shows that when \mathbb{N}_t is of the form (4), \mathcal{F}_n^{**} is in fact the optimal FVR for \mathbb{N}_t . This result was proved for $k=0$ in [24], [25], and for $k=1$ in [11]. In fact, once the basic properties of type classes are established, the proof should be rather insensitive to the order k , as it follows, essentially, from Lemma 1, from (3), and from the fact that for an arbitrary positive integer μ , the sum $\sum_{i=0}^m i c_i p^i$, subject to $\sum_{i=0}^m c_i p^i = \mu < p^{m+1}$, is maximized over vectors of nonnegative integers $[c_0, c_1, \dots, c_m]$ when c_0, c_1, \dots, c_m are the digits in the radix- p representation of μ (in our case, μ corresponds to the size of a type class).

Theorem 3: Let $\mathbb{N}_t = \{p^i \mid i \geq 0\}$ for some integer $p \geq 2$. Consider \mathcal{F}_n^{**} with target set \mathbb{N}_t , and let $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$ be

any FVR with the same target set. Then, for any n and any $P \in \mathcal{P}_k$, we have

$$L_P(\mathcal{M}) \leq L_P(\mathcal{M}^{**}).$$

Remark 5: The proposed variants of the Elias procedure assume knowledge of the (arbitrary) initial state s_0 . If the initial state is unknown (possibly nondeterministic), the procedure can consume k input symbols to synchronize its state and start at x_{k+1} , thus generating $E_P \log |T(X_{k+1}^n)|$ random bits (up to an additive constant), which is still asymptotically optimal. However, the pointwise optimality of Theorem 3 is lost. Nevertheless, the modified procedure can still be shown to be pointwise optimal in a more restrictive sense for the target sets covered by the theorem. Specifically, consider a setting in which an FVR is said to be universal if it is perfect for every $P \in \mathcal{P}_k$ and every initial state distribution (equivalently, for every fixed initial state). The definition of a type class in (3) is modified accordingly, and it is easy to see that the corresponding combinatorial characterization is given by the set of sequences with a given $\mathbf{n}(x^n)$ and fixed x_1^k . It can also be shown that, with the addition of $\alpha^k - 1$ free parameters (those corresponding to the distribution on the initial state), the type probabilities remain linearly independent, as required by the proof of Lemma 1. It follows that the modified FVR is optimal among FVRs that are perfect for every $P \in \mathcal{P}_k$ and every initial state distribution.

D. Twice-universal FVRs

In this subsection, we assume that the order k of the Markov source is not known, yet we want to produce a universal FVR whose model cost is not larger (up to third order terms) than the one we would incur had the value of k been given. To this end, as mentioned in Section I, we need to relax our requirement of a uniformly distributed output. This is necessary since, by Theorem 1 and (7), an FVR that is universal in \mathcal{P}_k would incur a model cost of the form $(K/2) \log n$, with $K = K(k) = \alpha^k(\alpha - 1)$, for any process in the class, including those of orders $k' < k$, which form a subclass of \mathcal{P}_k . However, for such a subclass, we aspire to achieve a smaller model cost proportional to $K(k')$.⁹ We assume throughout that \mathbb{N}_t is c -dense and that the fixed string determining the initial state is as long as needed (e.g., a semi-infinite all-zero string).

Let $Q_M(r)$ denote the output probability of $r \in [M]$, $M \in \mathbb{N}_t$, conditioned on $\mathcal{M}(x^n) = M$, for an FVR $\mathcal{F}_n = (\mathbb{N}_t, \rho, \mathcal{M})$. The distance of \mathcal{F}_n to uniformity is measured by

$$D(\mathcal{F}_n) \triangleq \sum_{M \in \mathbb{N}_t} \frac{P(\mathcal{M}(X^n) = M)}{M} \sum_{r, r' \in [M]} |Q_M(r) - Q_M(r')|. \quad (16)$$

⁹Of course, application of Procedure E2 with k replaced with a slowly growing function of n leads, for n sufficiently large, to a perfect FVR for any (fixed, but arbitrary) Markov order. However, the model cost incurred does not meet our efficiency demands.

For any distribution $R(\cdot)$ with support \mathcal{B} , we have

$$\begin{aligned} \sum_{x \in \mathcal{B}} \left| R(x) - \frac{1}{|\mathcal{B}|} \right| &= \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left| \sum_{y \in \mathcal{B}} (R(x) - R(y)) \right| \\ &\leq \frac{1}{|\mathcal{B}|} \sum_{x, y \in \mathcal{B}} |R(x) - R(y)|. \end{aligned}$$

In particular, the inner summation in (16) is lower-bounded by $M \sum_{r \in [M]} |Q_M(r) - 1/M|$. Therefore, our measure of uniformity is more demanding than the weighted L_1 measure used in [4]. Notice that, as in [4], the measure (16) is *unnormalized*. We aim at FVRs for which $D(\mathcal{F}_n)$ vanishes exponentially fast with n .

As in [14], our twice-universal FVR will rely on the existence of Markov order estimators with certain consistency properties, which are specified in Lemma 3 below. For concreteness, we will focus on a penalized maximum-likelihood estimator that, given a sample x^n from the source, chooses order $k(x^n)$ such that

$$k(x^n) = \arg \min_{k \geq 0} \left\{ \hat{H}_k(x^n) + \alpha^k \varphi(n) \right\} \quad (17)$$

where $\hat{H}_k(x^n)$ denotes the k -th order empirical conditional entropy for x^n , $\varphi(n)$ is a vanishing function of n , and ties are resolved, e.g., in favor of smaller orders. For example, $\varphi(n) = (\alpha - 1)(\log n)/(2n)$ corresponds to the asymptotic version of the MDL criterion [27]. The estimate $k(x^n)$ can be obtained in time that is linear in n by use of suffix trees [28], [29]. The set of n -tuples x^n such that $k(x^n) = i$ will be denoted \mathcal{A}_i^n . To state Lemma 3 we define, for a distribution $P \in \mathcal{P}_k$, the overestimation probability

$$P_{o/e}(n) \triangleq P(k(X^n) > k)$$

and, similarly, the underestimation probability

$$P_{u/e}(n) \triangleq P(k(X^n) < k).$$

Since the discussions in this subsection involve type classes of varying order, we will use the notation $T^{(k)}(x^n)$ to denote the type class of x^n with respect to \mathcal{P}_k .

Lemma 3 ([14]): For any $k \geq 0$ and any $P \in \mathcal{P}_k$, the estimator of (17) satisfies

- (a) $(n+1)^{\alpha^{k+1}} P_{o/e}(n)$ vanishes polynomially fast (uniformly in P and k) provided $\varphi(n) > \beta(\log n)/n$ for a sufficiently large constant β .
- (b) $P_{u/e}(n)$ vanishes exponentially fast.

Following [14], we consider a partition of \mathcal{A}^n in which the class of x^n , denoted $U(x^n)$, is given by

$$U(x^n) \triangleq T^{(k(x^n))}(x^n) \cap \mathcal{A}_{k(x^n)}^n. \quad (18)$$

Thus, two sequences are in the same class if and only if they estimate the same Markov order and are in the same type class with respect to the estimated order. Our twice-universal FVR, $\mathcal{F}_n^{(\text{TU})} = (\mathbb{N}_t, \rho^{(\text{TU})}, \mathcal{M}^{(\text{TU})})$, is given by replacing, in Procedure E2, $T^{(k)}(x^n)$ with $U(x^n)$ and $\mathcal{I}_T(x^n)$ with the index of x^n in an enumeration of $U(x^n)$.

Theorem 4: For $P \in \mathcal{P}_k$, the FVR $\mathcal{F}_n^{(\text{TU})}$ satisfies $D(\mathcal{F}_n^{(\text{TU})}) \leq 2P_{u/e}(n)$, and, for a suitable choice of $\varphi(n)$ in (17), its expected output length $L_P(\mathcal{M}^{(\text{TU})})$ satisfies

$$L_P(\mathcal{M}^{(\text{TU})}) - L_P(\mathcal{M}^*) = O(1) \quad (19)$$

provided \mathbb{N}_t is c -dense.

Remark 6: By Lemma 3, Theorem 4 states that the distance of $\mathcal{F}_n^{(\text{TU})}$ to uniformity is exponentially small whereas, by (6) and (7), its expected output length is essentially the same as that of \mathcal{F}_n^* . It should be pointed out, however, that Theorem 4 falls short of stating that the cost of twice-universality in terms of expected output length is asymptotically negligible. The reason is that, in principle, it could be the case that by allowing a small deviation from uniformity, as we do, we open the door for schemes that (with knowledge of k) produce an output significantly longer than \mathcal{F}_n^* . We conjecture that, just as in twice-universal simulation [14], this is not the case.

Remark 7: One problem in the implementation of $\mathcal{F}_n^{(\text{TU})}$ is that it requires a computationally efficient enumeration of $U(x^n)$. Such an enumeration appears to be elusive. Instead, the following FVR can be efficiently implemented: Compute $k(x^n)$ and apply Procedure E2 with $k=k(x^n)$. A variant of the proof of Theorem 4 shows that the output length of this scheme still satisfies (19), whereas its distance to uniformity is upper-bounded by $4[P_{u/e}(n) + P_{o/e}(n)]$. By Lemma 3, this means that a suitable choice of $\varphi(n)$ still guarantees vanishing distance, but we can no longer claim it to be exponentially small.

Proof of Theorem 4: Let \mathcal{U} denote the set of classes in the refinement of the partition (18) determined by Procedure E2 (see Remark 3), and let \mathcal{U}_M denote the subset of \mathcal{U} formed by classes of size $M \in \mathbb{N}_t$. For $U \in \mathcal{U}_M$, let $\rho_U^{-1}(r)$ denote the unique sequence in U such that $\rho^{(\text{TU})}(\rho_U^{-1}(r)) = r$, $r \in [M]$. Let $Q(r, M)$ denote the probability that $\mathcal{M}^{(\text{TU})}(x^n) = M$ and $\rho^{(\text{TU})}(x^n) = r$, $M \in \mathbb{N}_t$, $r \in [M]$, so that

$$Q_M(r) = \frac{Q(r, M)}{\sum_{j \in [M]} Q(j, M)} = \frac{Q(r, M)}{P(\mathcal{M}(X^n) = M)}.$$

Clearly,

$$Q(r, M) = \sum_{U \in \mathcal{U}_M} P(\rho_U^{-1}(r)).$$

By (16),

$$\begin{aligned} D(\mathcal{F}_n^{(\text{TU})}) &= \sum_{M \in \mathbb{N}_t} \frac{1}{M} \sum_{r, r' \in [M]} |Q(r, M) - Q(r', M)| \\ &\leq \sum_{M \in \mathbb{N}_t} \frac{1}{M} \sum_{U \in \mathcal{U}_M} \sum_{r, r' \in [M]} |P(\rho_U^{-1}(r)) - P(\rho_U^{-1}(r'))| \end{aligned}$$

which, given the existence of a one-to-one correspondence between $U \in \mathcal{U}_M$ and $[M]$, takes the form

$$D(\mathcal{F}_n^{(\text{TU})}) \leq \sum_{M \in \mathbb{N}_t} \frac{1}{M} \sum_{U \in \mathcal{U}_M} \sum_{u, v \in U} |P(u) - P(v)|. \quad (20)$$

Now, since U is a subset of a type class $T \in \mathcal{T}_n^{(k(x^n))}$, we have $P(u) = P(v)$ for all $u, v \in U$ whenever $k(x^n) \geq k$. In

addition, we have the following lemma, which is proved in Appendix C.

Lemma 4: For any distribution $R(\cdot)$ on a set containing \mathcal{B} , we have

$$\sum_{u,v \in \mathcal{B}} |R(u) - R(v)| \leq 2(|\mathcal{B}| - 1)R(\mathcal{B}).$$

Therefore, letting $\mathcal{U}_M^{u/e}$ denote the subset of \mathcal{U}_M formed by all the classes such that $k(x^n) < k$, (20) implies

$$D(\mathcal{F}_n^{(TV)}) \leq \sum_{M \in \mathbb{N}_t} \frac{2(M-1)}{M} \sum_{U \in \mathcal{U}_M^{u/e}} P(U) \leq 2P_{u/e}(n),$$

as claimed.

To lower-bound the expected output length, we first discard the output length produced by sequences which are not in \mathcal{A}_k^n . We then note that the claim of Theorem 2 is valid not only for expectations conditioned on a type (as implicit in its proof, see (13)), but also when conditioning on subsets of types, thus obtaining

$$L_P(\mathcal{M}^{(TV)}) \geq \sum_{T \in \mathcal{T}_n^{(k)}} P(T \cap \mathcal{A}_k^n) \log |T \cap \mathcal{A}_k^n| + O(1).$$

By [14, Lemma 1], the number of sequences in a type class T that estimate order k is $|T| - o(1)$ for suitable choices of $\varphi(n)$, provided that at least one sequence in T estimates order k (i.e., almost all the sequences in the type class estimate the right order). Therefore,

$$L_P(\mathcal{M}^{(TV)}) \geq \mathbf{E}_P \log |T_k(X^n)| - n [P_{u/e}(n) + P_{o/e}(n)] \log \alpha + O(1),$$

where we have also used the trivial bound $\log |T(x^n)| \leq n \log \alpha$ for sequences x^n outside \mathcal{A}_k^n . The claim then follows from Lemma 3. ■

IV. UNIVERSAL VARIABLE TO FIXED-LENGTH RNGS

A. Formal definition and preliminaries

We recall from Subsection II-A that a dictionary is a (possibly infinite) prefix-free set of finite strings $\mathcal{D} \subseteq \mathcal{A}^*$, which we assume full. A VFR is formally defined by a triplet $\mathcal{V} = (\mathcal{D}, \Phi, M)$ where \mathcal{D} is a dictionary, $M > 1$ is a fixed integer, and Φ is a function $\Phi : \mathcal{D} \rightarrow [M]$. For $N \geq 1$, the restriction to level N of \mathcal{D} is

$$\mathcal{D}_N = \{x^n \in \mathcal{D} \mid n \leq N\}.$$

Associated with \mathcal{D}_N is a *failure set* \mathcal{E}_N , defined as

$$\mathcal{E}_N = \{x^N \in \mathcal{A}^N \mid x^N \text{ has no prefix in } \mathcal{D}_N\}.$$

The strings in $\mathcal{D}_N \cup \mathcal{E}_N$ are identified with the leaves of a finite full tree, which is the truncation to depth N of the tree, $\mathbf{T}_{\mathcal{D}}$, corresponding to \mathcal{D} . Nevertheless, we will slightly abuse terminology, and refer to \mathcal{D}_N (alone) as a *truncated dictionary*. Notice that $\bigcup_{N \geq 1} \mathcal{E}_N$ corresponds to the set of internal nodes $\mathbf{I}_{\mathcal{D}}$ of $\mathbf{T}_{\mathcal{D}}$, whereas we recall that \mathcal{D} corresponds to the set of leaves of $\mathbf{T}_{\mathcal{D}}$.

The VFR \mathcal{V} generates random numbers from a process P by reading symbols from a realization of P until a string x^n in

\mathcal{D} is reached, at which point \mathcal{V} outputs $\Phi(x^n)$. The *truncated VFR* (TVFR) $\mathcal{V}_N = (\mathcal{D}_N, \Phi, M)$, operates similarly, except that it restricts the length of the input string to $n \leq N$, so that Φ is applied only to strings in \mathcal{D}_N , and the input may reach strings $x^N \in \mathcal{E}_N$, in which case the TVFR *fails* and outputs nothing.

A VFR $\mathcal{V} = (\mathcal{D}, \Phi, M)$ is *perfect* for $P \in \mathcal{P}_k$ if for every $n \geq 1$, either \mathcal{D}_n is empty, or $\Phi(X^n)$, conditioned on $X^n \in \mathcal{D}_n$, is uniformly distributed in $[M]$; \mathcal{V} is *universal* in \mathcal{P}_k if it is perfect for all $P \in \mathcal{P}_k$. By extension, we also refer to a TVFR that satisfies the same properties up to a certain length N as perfect or universal, as appropriate.

Next, we introduce tools that are instrumental in setting our objective. Let the dictionary \mathcal{D} satisfy $\sum_{x^* \in \mathcal{D}} P(x^*) = 1$ for all $P \in \mathcal{P}_k$. Notice that, if \mathcal{D} is finite, this condition is trivially satisfied by fullness; however, as discussed in Subsection II-A, it may not hold for a full infinite dictionary (for which the summation is understood as an infinite series in the usual manner), as it was shown in [16] that the Kraft inequality may be strict. Notice also that the condition is equivalent to $P(\mathcal{E}_N) \xrightarrow{N \rightarrow \infty} 0$. Let f be a real-valued function of $x^* \in \mathcal{A}^*$. The expectation of f over \mathcal{D} is denoted $\mathbf{E}_{P, \mathcal{D}} f(X^*)$ and, in case \mathcal{D} is infinite, it is given by

$$\mathbf{E}_{P, \mathcal{D}} f(X^*) = \lim_{N \rightarrow \infty} \sum_{x^* \in \mathcal{D}_N} P(x^*) f(x^*). \quad (21)$$

If f satisfies $0 \leq f(y^*) \leq f(x^*)$ for every prefix y^* of x^* (which is the case for functions such as string length or self-information), then it is easy to see that, due to the fullness of \mathcal{D} and to the vanishing failure probability, we have

$$\mathbf{E}_{P, \mathcal{D}} f(X^*) \geq \mathbf{E}_{P, \mathcal{D}_N \cup \mathcal{E}_N} f(X^*) \quad (22)$$

for every $N > 0$, provided the sequence on the right-hand side of (21) converges. But, since $\mathcal{D}_N \subseteq \mathcal{D}_N \cup \mathcal{E}_N$, the reverse inequality must hold when we let $N \rightarrow \infty$ on the right-hand side of (22). Therefore, the expectation also takes the form

$$\mathbf{E}_{P, \mathcal{D}} f(X^*) = \lim_{N \rightarrow \infty} \mathbf{E}_{P, \mathcal{D}_N \cup \mathcal{E}_N} f(X^*). \quad (23)$$

A useful tool in the analysis of $\mathbf{E}_{P, \mathcal{D}} f(X^*)$ is the so-called *leaf-average node-sum interchange theorem* (LANSIT) [30, Theorem 1], which states that

$$\begin{aligned} \mathbf{E}_{P, \mathcal{D}} f(X^*) &= \sum_{x^* \in \mathbf{I}_{\mathcal{D}}} P(x^*) \sum_{a \in \mathcal{A}} P(a|x^*) [f(x^*a) - f(x^*)] - f(\lambda). \end{aligned} \quad (24)$$

In particular, for $f(x^*) = |x^*|$, the LANSIT reduces to the well-known fact that the expected depth of the leaves of a tree equals the sum of the probabilities of its internal nodes. We will use the LANSIT also for $f(x^*) = 1/\log P(x^*)$ and $f(x^*) = n_s(x^*)$, $s \in \mathcal{A}^k$. The proof of the theorem, by induction on the number of nodes, is straightforward.

Consider a VFR \mathcal{V} that is perfect for P . The quantity

$$\begin{aligned} L_P(\mathcal{D}_N) &\triangleq \mathbf{E}_{P, \mathcal{D}_N \cup \mathcal{E}_N} |X^*| \\ &= \sum_{n=1}^N \sum_{x^n \in \mathcal{D}_N} nP(x^n) + NP(\mathcal{E}_N) \end{aligned} \quad (25)$$

is an appropriate figure of merit for \mathcal{V} at truncation level N , as it measures the *expected input length*, namely the amount of “raw” random data that the VFR consumes in order to produce a perfectly uniform distribution on $[M]$, when restricted to inputs of length at most N . The expected input length includes the cost of “unproductive” input that is consumed when the truncated VFR fails (second term on the rightmost side of (25)). The figure of merit for \mathcal{V} is given by

$$L_P(\mathcal{D}) = \lim_{N \rightarrow \infty} L_P(\mathcal{D}_N) \quad (26)$$

which, by (23), coincides with the expected dictionary length provided $\sum_{x^* \in \mathcal{D}} P(x^*) = 1$.¹⁰

We are interested in universal VFRs that minimize these measures simultaneously for all $P \in \mathcal{P}_k$, either in a pointwise sense, i.e., minimizing $L_P(\mathcal{D}_N)$ for all N , or asymptotically, i.e., minimizing the limit $L_P(\mathcal{D})$. A secondary objective is to minimize the failure probability $P(\mathcal{E}_N)$. Finally, we are interested in *computationally efficient implementations*, namely, VFR procedures that process the input sequentially, and run in time polynomial in the consumed input length.

By the LANSIT, we have

$$L_P(\mathcal{D}_N) = \sum_{x^* \in \mathcal{I}_{\mathcal{D}_N} \cup \mathcal{E}_N} P(x^*) = \sum_{n=0}^{N-1} P(\mathcal{E}_n). \quad (27)$$

Therefore, the limit in (26) exists if and only if $P(\mathcal{E}_n)$ is summable. We will show that, in fact, the failure probability in our constructions vanishes exponentially fast, so that the limit does exist (and equals the expected dictionary length).

In the sequel, we will make extensive use of the following notation: For $T \in \mathcal{T}_n$ and $\mathcal{S} \subset \mathcal{A}^*$, $\mathcal{S}(T) \triangleq \mathcal{S} \cap T$ (this definition is extended to the case in which \mathcal{S} is a set of nodes in a tree). In particular, we have $\mathcal{I}_{\mathcal{D}}(T) = \mathcal{E}_n(T)$ and $\mathcal{T}_{\mathcal{D}}(T) = \mathcal{D}_n(T) \cup \mathcal{E}_n(T)$.

B. Necessary and sufficient condition for universality of VFRs

The analog of Lemma 1 for VFRs is given in the following lemma. The proof is similar, and is presented, for completeness, in Appendix D.

Lemma 5: Let $\mathcal{V} = (\mathcal{D}, \Phi, M)$ be a VFR satisfying the following condition: For every n and every $T \in \mathcal{T}_n$, the number of sequences $x^n \in \mathcal{D}(T)$ such that $\Phi(x^n) = r$ is the same for all $r \in [M]$ (in particular, $|\mathcal{D}(T)|$ is a multiple of M). Then, \mathcal{V} is universal in \mathcal{P}_k . If \mathcal{V} does not satisfy the condition, then it can only be perfect for processes P with parameter \mathbf{p} in a subset Ψ'_0 of measure zero in Ψ .

An analog of Corollary 1 for the VFR case is also straightforward. Notice that if $|\mathcal{D}(T)|$ is a multiple of M , then it is trivial to define Φ so that \mathcal{V} satisfies the condition of the lemma. Therefore, designing a universal VFR is essentially equivalent to designing a dictionary \mathcal{D} such that

$$|\mathcal{D}(T)| = j_T M, \quad \forall T \in \mathcal{T}_n, \quad \forall n \in \mathbb{N}^+, \quad (28)$$

¹⁰If, instead, $\sum_{x^* \in \mathcal{D}} P(x^*) < 1$, it may be the case that the expected dictionary length converges (again, [16] provides an example of such a tree) while, clearly, the expected input length diverges. In this case, the expected dictionary length is of no interest since, with a positive probability, an input sample will not have a prefix in the dictionary (i.e., the VFR will not stop).

where j_T is a nonnegative integer dependent on T . In fact, in our discussions, we will focus on the condition (28) and assume that a suitable mapping Φ is defined when the condition holds.

Remark 8: Lemma 5 implies that our universal VFRs are akin to the *even procedures* discussed in [5] and [6] (the term *even* derives from the fact that the emphasis in [5] is on the case $M = 2$, although the more general case is also mentioned). In our case, the necessity of the condition (28) stems from our requirement that the VFR be perfect at every truncation level N . When this requirement is relaxed, the condition need no longer hold, as evidenced by some of the procedures presented in [5] and [6]. As we will see, such a relaxation may reduce the expected input length of the optimal VFR only by a negligible amount relative to the main asymptotic term (see also Example 1 below).

Remark 9: Notice that the condition on universality in Lemma 5 depends only on the *sizes* of the sets $\mathcal{D}(T)$, but not on their composition. Clearly, the same holds for the expected length and the failure probability of a (truncated) dictionary, since sequences of the same type have the same length and probability. We conclude that the main properties of interest for a VFR are fully determined by the *type profile* of its dictionary, namely, the sequence of numbers $\{|\mathcal{D}(T)|\}_{T \in \mathcal{T}_n, n \geq 1}$.

Define

$$N_0(M) = \min \{ n \mid \exists T \in \mathcal{T}_n \text{ such that } |T| \geq M \}. \quad (29)$$

An immediate consequence of Lemma 5 is that if \mathcal{D} is the dictionary of a universal VFR, then $n \geq N_0(M) \geq (\log M)/(\log \alpha)$ for every $x^n \in \mathcal{D}$, where the second inequality follows from (29) and $|T| \leq \alpha^n$.

C. Optimality of a “greedy” universal VFR

We describe the (conceptual) construction of a universal VFR, and prove its optimality. The construction is “greedy” in the sense that, at every point, it tries to add to the dictionary as many sequences of a given length as allowed by the necessary condition of Lemma 5, and by the prefix condition. In this sense, the procedure can be seen as a counterpart, for VFRs, to Elias’s scheme for FVRs (recall the discussion on the “greediness” of Procedure E2 in Subsection III-C). As in the FVR case, it will turn out that greediness pays off, and the constructed VFR will be shown to be optimal in a pointwise, non-asymptotic sense. The difficulty in establishing this optimality will reside in the fact that sequences that get included in \mathcal{D} “block” all of their continuations from membership in \mathcal{D} . It seems possible, in principle, that it might pay off not to include some sequences of a given length, even though the conditions governing the construction allowed their inclusion, so as to increase our choices for longer sequences. We will prove that, in fact, this is not the case.

Procedure G1 in Fig. 2 shows the construction of a TVFR $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi, M)$. The VFR $\mathcal{V}^* = (\mathcal{D}^*, \Phi, M)$ is then obtained by letting $\mathcal{D}^* = \bigcup_{N \geq 1} \mathcal{D}_N^*$. The procedure starts from an empty dictionary, and adds to it sequences of increasing length $n = 1, 2, 3, \dots$, sequentially, so that for each type class $T \in \mathcal{T}_n$, it “greedily” augments \mathcal{D}^* with the largest possible

Input: Integers $M \geq 2$, $N \geq 1$.
Output: TVFR $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi^*, M)$.

- 1) Set $n = 1$, $\mathcal{D}_N^* = \emptyset$, $\mathcal{E}_N^* = \mathcal{A}$.
 - 2) For each type class $T \in \mathcal{T}_n$, do:
 - a) Let $j_T = \lfloor |\mathcal{E}_N^*(T)|/M \rfloor$. Select any subset of $j_T M$ sequences from $\mathcal{E}_N^*(T)$, add them to \mathcal{D}_N^* , and remove them from $\mathcal{E}_N^*(T)$.
 - b) Let $\mathcal{I}(y^n)$ denote the index of $y^n \in \mathcal{D}_N^*(T)$ in some ordering of $\mathcal{D}_N^*(T)$. Define
$$\Phi^*(y^n) = \mathcal{I}(y^n) \bmod M, \quad y^n \in \mathcal{D}_N^*(T).$$
 - 3) If $n = N$, **stop**. Otherwise, for each sequence $x^n \in \mathcal{E}_N^*$, remove x^n and add all the sequences in $\{x^n a, a \in \mathcal{A}\}$ to \mathcal{E}_N^* . Set $n \leftarrow n + 1$ and go to Step 2.
-

Fig. 2. Procedure G1: Greedy TVFR construction.

number of sequences in T that is a multiple of M and such that these sequences have no prefix in \mathcal{D}^* . The procedure is presented as a characterization of \mathcal{V}^* , rather than as a computational device. An effective, sequential implementation of \mathcal{V}^* will be presented in Subsection IV-E.

Theorem 5: The TVFR $\mathcal{V}_N^* = (\mathcal{D}_N^*, \Phi^*, M)$ constructed by Procedure G1 is universal.

Proof: The fact that the set $\mathcal{D}_N^* \cup \mathcal{E}_N^*$ constructed by the procedure is prefix-free and full can be easily seen by induction in n : the procedure starts with $\mathcal{D}_N^* \cup \mathcal{E}_N^* = \mathcal{A}$ and at each iteration it moves sequences from \mathcal{E}_N^* to \mathcal{D}_N^* (Step 2a) and replaces the remaining sequences of length n in \mathcal{E}_N^* with a full complement of children of length $n+1$ (Step 3). Sequences from a type class T are added to \mathcal{D}_N^* in sets of size $j_T M$, for some $j_T \geq 0$ depending on T (Step 2a), and are assigned uniformly to symbols in $[M]$ (Step 2b). Therefore, by Lemma 5, the constructed TVFR is universal. ■

The following key lemma is the basis for the proof of pointwise optimality of \mathcal{V}_N^* .

Lemma 6: Let \mathcal{D} be the dictionary of a universal VFR. Then, for every type class $T \in \mathcal{T}_n$, we have

$$|\mathcal{E}_n(T)| \equiv |T| \pmod{M} \quad (30)$$

and, in particular, for the dictionary \mathcal{D}^* , we have

$$|\mathcal{E}_n^*(T)| = |T| \bmod M. \quad (31)$$

Moreover, if $|\mathcal{E}_n(T)| < M$ for every type class $T \in \mathcal{T}_n$, $1 \leq n \leq N$, then $|\mathcal{D}(T)| = |\mathcal{D}^*(T)|$ for all $T \in \mathcal{T}_n$, $1 \leq n \leq N$.

Proof: Let $T \in \mathcal{T}_n$ and $T' \in \mathcal{T}_m$, with $m < n$. For an arbitrary $y^m \in T'$, consider the set

$$\Delta(T, T') = \{z^{n-m} \in \mathcal{A}^{n-m} \mid y^m z^{n-m} \in T\} \quad (32)$$

which, by Fact 1, depends only on T and T' and is independent of the choice of y^m (in fact, $\Delta(T, T') \in \mathcal{T}_{n-m}$, but with an initial state equal to the common final state of the sequences in T'). Now, since \mathcal{D} is a full prefix set, if $y^m \in \mathcal{D}(T')$ then $y^m z^{n-m} \notin \mathbf{T}_{\mathcal{D}}(T)$ and, conversely, each $x^n \in T \setminus \mathbf{T}_{\mathcal{D}}(T)$ must have a unique proper prefix $x^m \in \mathcal{D}$, in a type class $T' \in \mathcal{T}_m$. Since a sequence in T either has a proper prefix in \mathcal{D} or it corresponds to a node in $\mathbf{T}_{\mathcal{D}}$, in which case the node

is either a leaf (sequences in $\mathcal{D}(T)$) or internal (sequences in $\mathcal{E}_n(T)$), it follows that

$$|\mathcal{E}_n(T)| = |T| - |\mathcal{D}(T)| - \sum_{m=1}^{n-1} \sum_{T' \in \mathcal{T}_m} |\Delta(T, T')| \cdot |\mathcal{D}(T')|, \quad (33)$$

where the double summation is the number of sequences in T that have a proper prefix in \mathcal{D} . By Lemma 5, each $|\mathcal{D}(T')|$ in (33), as well as $|\mathcal{D}(T)|$, must be divisible by M , implying (30). Equation (31) then follows from the construction in Procedure G1, which guarantees that $|\mathcal{E}_n^*(T)| < M$.

Next, assume $|\mathcal{D}(T)| \neq |\mathcal{D}^*(T)|$ for some $T \in \mathcal{T}_n$, $1 \leq n \leq N$. Without loss of generality, assume n is the smallest such integer, so that $|\mathcal{D}(T')| = |\mathcal{D}^*(T')|$ for all $T' \in \mathcal{T}_m$, $m < n$. By (33), we have $\mathcal{E}_n(T) \neq \mathcal{E}_n^*(T)$. But, if $|\mathcal{E}_n(T)| < M$, by (30) and (31), we must have $|\mathcal{E}_n(T)| = |\mathcal{E}_n^*(T)|$. Therefore, we must also have $|\mathcal{D}(T)| = |\mathcal{D}^*(T)|$ for every $T \in \mathcal{T}_n$, $1 \leq n \leq N$. ■

The following theorem establishes the pointwise optimality of \mathcal{V}_N^* and also the uniqueness of the optimal type profile for a universal VFR.

Theorem 6: Let $\mathcal{V} = (\mathcal{D}, \Phi, M)$ be a universal VFR. Then, for every $N \geq 0$, we have $L_P(\mathcal{D}_N^*) \leq L_P(\mathcal{D}_N)$ and $P(\mathcal{E}_N^*) \leq P(\mathcal{E}_N)$ for all $P \in \mathcal{P}_k$. Moreover, if $|\mathcal{D}(T)| \neq |\mathcal{D}^*(T)|$ for any n and $T \in \mathcal{T}_n$, then $L_P(\mathcal{D}_N^*) < L_P(\mathcal{D}_N)$ for all $N > n$ and all $P \in \mathcal{P}_k$.

Proof: If \mathcal{V} is universal, then, by Lemma 6, we have $|\mathcal{E}_n^*(T)| \leq |\mathcal{E}_n(T)|$ for all $T \in \mathcal{T}_n$ and thus, since sequences of the same type are equiprobable, $P(\mathcal{E}_N^*) \leq P(\mathcal{E}_N)$ for every $N \geq 0$. Moreover, by (27), we also have $L_P(\mathcal{D}_N^*) \leq L_P(\mathcal{D}_N)$. Now, if there exists $T \in \mathcal{T}_n$ such that $|\mathcal{D}(T)| \neq |\mathcal{D}^*(T)|$ then, by Lemma 6, we have $|\mathcal{E}_{n'}^*(T')| \geq M$ for some $T' \in \mathcal{T}_{n'}$, with $n' \leq n$. Therefore, $P(\mathcal{E}_{n'}^*) < P(\mathcal{E}_{n'})$ which, by (27), implies that $L_P(\mathcal{D}_N^*) < L_P(\mathcal{D}_N)$ for all $N > n$ and all $P \in \mathcal{P}_k$. ■

Remark 10: A modification analogous to the one presented in Remark 5 is valid in the VFR setting as well for the case in which the initial state is not deterministic. Specifically, the VFR consumes k input symbols and then applies \mathcal{V}^* with initial state x_1^k . Equivalently, the dictionary of the modified procedure corresponds to a tree obtained by taking a balanced tree of depth k , and “hanging” from each leaf s the tree corresponding to \mathcal{V}^* with initial state s . Again, this VFR is optimal among VFRs that are perfect for every $P \in \mathcal{P}_k$ and every initial state distribution.

By (27) and (31) in Lemma 6, the expected dictionary length of \mathcal{V}_N^* is given by

$$L_P(\mathcal{D}_N^*) = \sum_{n=0}^{N-1} \sum_{T \in \mathcal{T}_n} \frac{|T| \bmod M}{|T|} P(T). \quad (34)$$

As the exact formula in (34) appears to provide little insight into the performance of \mathcal{V}_N^* in general (in terms of both expected dictionary length and failure probability, except for special cases such as in Example 1 below), a precise estimate is deferred to subsections IV-F and IV-G. In particular, it will be shown that $P(\mathcal{E}_N^*)$ decays exponentially fast with N and, consequently, as discussed in Subsection IV-A, $L_P(\mathcal{D}_N^*)$ converges to the expected dictionary length of \mathcal{D}^* , which is optimal among all universal VFRs with vanishing failure probability.

Example 1: Consider the VFR \mathcal{V}^* for the Bernoulli class and $M = 3$. Clearly, $N_0(3) = 3$, and there exist two type classes of size 3 in \mathcal{T}_3 , namely $T = \{001, 010, 100\}$ and $T' = \{011, 101, 110\}$. By Procedure G1, both T and T' are included in \mathcal{D}^* and, thus, $\mathcal{E}_3^* = \{000, 111\}$. Next, the procedure considers the set of extensions of 000 and 111. This set does not contain a subset of size 3 of sequences of the same type for any $n < 6$. For $n = 6$, four such subsets do exist, namely, the concatenations of 000 and 111 with the sequences in T and T' . Consequently, $\mathcal{E}_6^* = \{000000, 000111, 111000, 111111\}$ (notice that $|T(000111)| = 20$ and there are two sequences of this type in \mathcal{E}_6^* , as predicted by (31) in Lemma 6). Again, it can be readily verified that the set of extensions of the sequences in \mathcal{E}_6^* does not contain a subset of size 3 of sequences of the same type for $n = 7$, but such subsets do exist for $n = 8$ (e.g., $\{00011101, 00011110, 11100001\}$). The construction proceeds in a similar fashion.

Next, we let $P(0) = p = 1 - q$ and bound $L_P(\mathcal{D}^*)$ in the example. By (34), we have

$$L_P(\mathcal{D}^*) = \sum_{n=0}^{\infty} \sum_{k=0}^n \binom{n}{k} \text{mod } 3 \, p^k q^{n-k}.$$

To derive a lower bound, we take the terms in the sum corresponding to $k=0, 1, 2$ and $n-k=0, 1, 2$, and solve the resulting sums for n (the cases $k=0, 1$ and their complements are straightforward, whereas for $k=2$ we note that $\binom{n}{2} = 1 \text{ mod } 3$ when $n = 2 \text{ mod } 3$, and $\binom{n}{2} = 0 \text{ mod } 3$ otherwise).¹¹ After tedious but straightforward computations, it can be shown that

$$L_P(\mathcal{D}^*) > \frac{1}{pq} - \frac{pq(1+3pq)}{3+p^2q^2} \geq \frac{1}{pq} - \frac{1}{7}. \quad (35)$$

Notice that a direct generalization of von Neumann's scheme [1] for the case $M = 3$ would proceed as Procedure G1 up to $n = 3$, but, in case $x^3 \in \mathcal{E}_3^*$, it would iterate the procedure "from scratch" until an output is produced. This scheme is clearly universal and it is straightforward to show that its expected dictionary length is $1/(pq)$ which, therefore, by Theorem 6, is an upper bound on $L_P(\mathcal{D}^*)$. However, consider the following variant of the iterated scheme: the VFR, denoted $\bar{\mathcal{V}} = (\bar{\mathcal{D}}, \bar{\Phi}, 3)$, outputs $\bar{\Phi}(01) = 0$, $\bar{\Phi}(10) = 1$, and $\bar{\Phi}(001) = \bar{\Phi}(110) = 2$, whereas on inputs 000 and 111, the procedure is iterated. Since $P(01) = P(10) = P(001) + P(110) = pq$, $\bar{\Phi}$ is uniformly distributed for all values of p . Also, since the expected dictionary length for the first iteration is $3 - 2pq$, and an iteration occurs with probability $p^3 + q^3$, overall, we have

$$L_P(\bar{\mathcal{D}}) = \frac{3 - 2pq}{1 - p^3 - q^3} = \frac{1}{pq} - \frac{2}{3} < L_P(\mathcal{D}^*)$$

where the inequality follows from (35). The reason $\bar{\mathcal{V}}$ can outperform \mathcal{V}^* is that it *does not* preserve universality under truncation (note that the condition of Lemma 5 is not satisfied for $\bar{\mathcal{V}}$, and the truncated scheme $\bar{\mathcal{V}}_N$ is not perfect whenever $N \equiv 2 \pmod{3}$). Perfect VFRs without the truncation requirement are studied in [5], [6]: in particular, it is shown in [6]

¹¹More terms can be estimated using Lucas's Theorem, which applies to any prime M .

that no single VFR can be optimal, in this sense, for all values of p . Notice also that using the upper bound on $L_P(\mathcal{D}^*)$ we obtain $L_P(\mathcal{D}^*) - L_P(\bar{\mathcal{D}}) < 2/3$. In fact, as will be discussed in Section IV-G, as M grows, the cost of requiring perfection under truncation becomes negligible. \square

The performance analysis in subsections IV-F and IV-G, as well as the design of an efficient implementation in Subsection IV-E, require the discussion of additional properties of type classes, dictionaries, and their interactions. We present these properties in the next subsection.

D. Additional properties

We next describe a decomposition of a type class $T \in \mathcal{T}_n$. For $u \in \mathcal{A}^\ell$, $0 \leq \ell \leq n$, let

$$S_u(T) = \{x^n \mid x^n \in T, x_{n-k-\ell+1}^{n-k} = u\} \quad (36)$$

(by our assumptions on initial states, and by the range defined for ℓ , $S_u(T)$ is well defined even if some of the indices of $x_{n-k-\ell+1}^{n-k}$ in (36) are negative). Clearly, we can decompose T as

$$T = \bigcup_{u \in \mathcal{A}^\ell} S_u(T), \quad (37)$$

where, by (36) and Fact 1, the sequences in $S_u(T)$ coincide in their last $k + \ell$ symbols. For $u \in \mathcal{A}^\ell$, define

$$S_u^-(T) = \{x^{n-\ell} \mid x^n \in S_u(T)\}, \quad 0 \leq \ell \leq n, \quad (38)$$

and $S_u^-(T) = \emptyset$, $\ell > n$. From the definitions (36), (38), it is readily verified that, for $u, v \in \mathcal{A}^*$,

$$S_{vu}^-(T) = S_v^-(S_u^-(T)). \quad (39)$$

What makes the sets $S_u^-(T)$ useful is the fact that they are, generally, type classes themselves, as established in the following lemma.

Lemma 7: If $S_u^-(T)$ is not empty then $S_u^-(T) \in \mathcal{T}_{n-\ell}$.

Proof: We prove the result by induction on ℓ . For $\ell = 0$, the claim is trivial, since $S_\lambda^-(T) = T$. Assume the claim is true for all ℓ' , $0 \leq \ell' < \ell$, and consider a string $u = au'$, $a \in \mathcal{A}$, $u' \in \mathcal{A}^{\ell-1}$. If $S_u^-(T)$ is not empty, then neither is $S_{u'}^-(T)$, and, by the induction hypothesis, we have $S_{u'}^-(T) = T' \in \mathcal{T}_m$, where $m = n - \ell + 1$. Consider a sequence $x^m \in T'$. The type of sequences in $S_u^-(T)$ differs from that of x^m in one count of x_m , which is deducted from $n_s^{(x_m)}(x^m)$, $s = ax_{m-k+1}^{m-1}$, if $k > 0$, or from the global count of $x_m = a$ if $k = 0$. In either case, by Fact 1 and the definition of $S_u^-(T)$, both s and x_m are invariant over $S_u^-(T)$, and, therefore, $S_u^-(T) \subseteq T''$ for some type class $T'' \in \mathcal{T}_{m-1}$. On the other hand, if a sequence $y^{m-1} \in T''$ is extended with a symbol y_m (whether $y_m = a$ when $k = 0$ or y_m is the invariant final symbol x_m of sequences in T' when $k > 0$), then the counts of T' are restored, so we have $y^m \in S_a(T')$, and, hence, $T'' \subseteq S_a^-(T') = S_u^-(T)$, where the last equality follows from (38). Therefore, $S_u^-(T) = T''$. \blacksquare

Remark 11: When $\ell > k + 1$, the type classes $S_u^-(T)$ and $S_{u'}^-(T)$ may coincide even if $u \neq u'$. Specifically, letting $s_f \in \mathcal{A}^k$ denote the final state of T , this situation arises if and only if $u = u_1^k v$, $u' = u_1^k v'$, and $T(vs_f) = T(v's_f)$, where both type classes assume an initial state u_1^k (it is easy

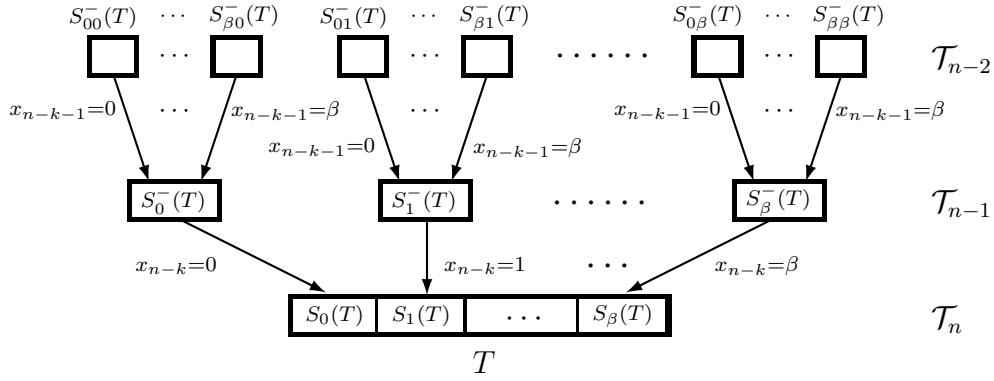


Fig. 3. Type class relations (with $\mathcal{A} = \{0, 1, \dots, \beta\}$, $\beta = \alpha - 1$).

to see that this situation requires $|v| > 1$). In fact, the type class $T(vsf)$ is precisely the set $\Delta(T, T')$ defined in (32), for $T' = S_u^-(T)$.

Equations (37)–(38) and Lemma 7 show how we can trace the origins of sequences in a type class T to the type classes $S_u^-(T)$ of their prefixes. This relation between type classes is illustrated, for $\ell = 2$, in Fig. 3. Extending the figure recursively, using (39), T can be seen as being at the root of an α -ary tree tracking the path, through shorter type classes, of sequences that end up in T . This structure will be useful in the derivation of various results in the sequel.

We now apply the foregoing type class relations to obtain a recursive characterization of $\mathcal{E}_n(T)$, $T \in \mathcal{T}_n$, for a given dictionary \mathcal{D} .

Lemma 8: For any dictionary \mathcal{D} and any class type $T \in \mathcal{T}_n$, there is a one-to-one correspondence between $\mathbf{T}_{\mathcal{D}}(T)$ and $\bigcup_{a \in \mathcal{A}} \mathbf{I}_{\mathcal{D}}(S_a^-(T))$, which implies

$$|\mathcal{E}_n(T)| = \sum_{a \in \mathcal{A}} |\mathcal{E}_{n-1}(S_a^-(T))| - |\mathcal{D}(T)|, \quad (40)$$

where we take $\mathcal{E}_{n-1}(\emptyset) = \emptyset$ (some of the sets $S_a^-(T)$ may be empty).

Proof: Clearly, by the decomposition (37), each node in $\mathbf{T}_{\mathcal{D}}(T)$ is a child of a node in $\mathbf{I}_{\mathcal{D}}(S_a^-(T))$ for some $a \in \mathcal{A}$. Conversely, a node in $\mathbf{I}_{\mathcal{D}}(S_a^-(T))$ has a *unique* child in $\mathbf{T}_{\mathcal{D}}(T)$, in the direction prescribed by the decomposition. To complete the proof, we recall that, for any $T' \in \mathcal{T}_{n'}$ and any n' , the nodes in $\mathbf{I}_{\mathcal{D}}(T')$ correspond to the sequences in $\mathcal{E}_{n'}(T')$, whereas the sequences in $\mathcal{D}(T')$ correspond to the leaves in $\mathbf{T}_{\mathcal{D}}(T')$. ■

E. Sequential implementation of \mathcal{V}^*

Procedure G1 in Fig. 2 constructs dictionaries of depth N , for arbitrarily large values of N . In a practical VFR application, however, dictionaries are not actually constructed. Instead, what is required is a procedure that reads the input sequence $x_1 x_2 \dots x_n \dots$, sequentially, and, for each n , makes a decision as to whether it should produce an output (and what that output should be) and stop, or continue processing more input. Procedure G2 in Fig. 4 describes such a sequential implementation of \mathcal{V}^* , without truncation. The procedure can

Input: Sequence $x_1 x_2 x_3 \dots x_n \dots$, integer $M > 1$.
Output: Number $r \in [M]$.

- 1) Set $\mathcal{I}_{\mathcal{E}} = 0$, $n = 0$, $\mathbf{n}(x^n) = \mathbf{0}$.
 - 2) Increment n , read x_n and update $\mathbf{n}(x^n)$. Let $T = T(x^n)$.
 - 3) Compute $|\mathcal{E}_{n-1}(S_a^-(T))| = |\mathcal{E}_{n-1}(S_a^-(T))| \bmod M$, for each $a \in \mathcal{A}$.
 - 4) Set $\mathcal{I}_{\mathbf{T}_{\mathcal{D}}} = \sum_{a < x_{n-k}} |\mathcal{E}_{n-1}(S_a^-(T))| + \mathcal{I}_{\mathcal{E}}$.
 - 5) Set $j_T = \left\lfloor \sum_{a \in \mathcal{A}} |\mathcal{E}_{n-1}(S_a^-(T))| / M \right\rfloor$.
 - 6) If $\mathcal{I}_{\mathbf{T}_{\mathcal{D}}} < j_T M$ then **output** $r = \mathcal{I}_{\mathbf{T}_{\mathcal{D}}} \bmod M$ and **stop**. Otherwise, set $\mathcal{I}_{\mathcal{E}} = \mathcal{I}_{\mathbf{T}_{\mathcal{D}}} - j_T M$ and **go to** Step 2.
-

Fig. 4. Procedure G2: Sequential implementation of \mathcal{V}_N^* .

easily be modified to implement a TVFR for arbitrary N , with a possible failure exit.

The procedure relies on a sequential alphabetic enumeration of $\mathbf{T}_{\mathcal{D}}(T(x^n))$, which defines a partition of this set into $\mathcal{E}_n(T(x^n))$ and $\mathcal{D}(T(x^n))$, and yields an enumeration of the two parts. These enumerations are based, in turn, on the one-to-one correspondence of Lemma 8, and determine whether an output is produced, and the value of the output. We assume a total (alphabetic) order $<$ of the elements of \mathcal{A} ; for the purpose of comparing sequences of length n , significance increases with the coordinate index (i.e., x_n is the most, and x_1 the least, significant symbol in x^n). We assume, recursively, that after processing x^{n-1} we have its index $\mathcal{I}_{\mathcal{E}}(x^{n-1})$ in $\mathcal{E}_{n-1}(S_{x_{n-k}}^-(T(x^n)))$ (notice that $T(x^{n-1}) = S_{x_{n-k}}^-(T(x^n))$). In Fig. 4, this index is assumed stored in the variable $\mathcal{I}_{\mathcal{E}}$ when Step 2 is reached. Since all the sequences in $T = T(x^n)$ coincide in their last k symbols, if $y^n \in T$ and $y_{n-k} < x_{n-k}$, then $y^n < x^n$ in the alphabetical order. Therefore, using the one-to-one correspondence of Lemma 8, the index of x^n in $\mathbf{T}_{\mathcal{D}}(T)$ is given by

$$\mathcal{I}_{\mathbf{T}_{\mathcal{D}}}(x^n) = \sum_{a < x_{n-k}} |\mathcal{E}_{n-1}(S_a^-(T))| + \mathcal{I}_{\mathcal{E}}(x^{n-1}). \quad (41)$$

In Fig. 4, this computation is performed in Step 4, based on the value of the aforementioned index $\mathcal{I}_{\mathcal{E}}(x^{n-1})$ available when Step 2 was reached, and on the values $|\mathcal{E}_{n-1}(S_a^-(T))|$ computed in Step 3. The computations in Step 3, in turn, can be derived from (31), by means of Whittle's formula [17]

applied to $S_a^-(T) \in \mathcal{T}_{n-1}$. Notice that the type $\mathbf{n}^{(a)}$ associated with $S_a^-(T)$, which is required to evaluate Whittle's formula, is easily obtained from the type $\mathbf{n}(x^n)$. In Step 5, the factor j_T of $|\mathcal{D}(T)| = j_T M$ is obtained, based again on the quantities computed in Step 3 and on (40). We assume that $\mathcal{D}(T)$ consists of the first $j_T M$ sequences in the alphabetic ordering of $\mathbf{T}_{\mathcal{D}}(T)$. If the index (41) of x^n in this ordering is less than $j_T M$, as checked in Step 6, then x^n is in the dictionary, and an output is produced. Setting the output value as in Step 6 guarantees that sequences in $\mathcal{D}(T)$ are assigned uniformly to values in $[M]$, as required by the condition of Lemma 5. If x^n is not in \mathcal{D} , then its index in $\mathcal{E}_n(T)$ is obtained by subtracting $|\mathcal{D}(T)| = j_T M$ from its index in $\mathbf{T}_{\mathcal{D}}(T)$, so $\mathcal{E}_n(T)$ inherits the alphabetic ordering of $\mathbf{T}_{\mathcal{D}}(T) = \mathcal{D}(T) \cup \mathcal{E}_n(T)$, and the assumptions for the next iteration are satisfied. Since Whittle's formula can be evaluated in time polynomial in n , the procedure in Fig. 4 runs in polynomial time.

F. Performance: Preliminaries

We study the performance of \mathcal{V}_N^* in terms of expected dictionary length and failure probability for large N . First, we show that the failure probability vanishes exponentially fast and that, as a result, the expected dictionary length converges. We then characterize the asymptotic behavior with respect to M of the convergence value, up to an additive constant independent of M . To this end, for sufficiently large N , we derive a lower bound on $L_P(\mathcal{D}_N)$ for any truncated dictionary \mathcal{D}_N derived from a universal VFR and we show that the bound is achievable within such a constant. For the achievability result we will not use the optimal universal VFR \mathcal{V}^* , but a different VFR, for which the analysis is simpler.¹²

Theorem 7: For every $P \in \mathcal{P}_k$, $P(\mathcal{E}_N^*)$ decays exponentially fast with N .

Proof: By (31), recalling that sequences of the same type are equiprobable, for any $\epsilon > 0$ we have

$$P(\mathcal{E}_N^*) < P\left(|T(X^N)| < 2^{N(\bar{H}_P(X) - \epsilon)}\right) + M2^{-N(\bar{H}_P(X) - \epsilon)}. \quad (42)$$

Now, recalling that, if the maximum-likelihood estimator for x^N is bounded away from the boundary of Ψ , then $|T(x^N)|$ is exponential in $N\hat{H}_k(x^N)$ for large N , the event $|T(X^N)| < 2^{N(\bar{H}_P(X) - \epsilon)}$ is a large deviations one. Therefore, both terms on the rightmost side of (42) decay exponentially fast with N . ■

As argued, by (27), Theorem 7 implies that $L_P(\mathcal{D}_N^*)$ converges to the expected dictionary length of \mathcal{D}^* . Next, we develop the basic tools we will use in the characterization of $L_P(\mathcal{D}_N^*)$, starting with some definitions. Throughout this subsection we assume that all dictionaries satisfy that, for every $P \in \mathcal{P}_k$, $P(\mathcal{E}_N)$ vanishes as $N \rightarrow \infty$. In the next subsection, as we apply these tools to specific dictionaries, this property will need to be verified in each case.

The entropy of a dictionary \mathcal{D} is defined as $H_P(\mathcal{D}) \triangleq -\mathbf{E}_{P, \mathcal{D}} \log P(X^*)$. When the initial state s differs from s_0 , we

use the notation $H_{P,s}(\mathcal{D})$ for this entropy. As shown in (23), due to the vanishing failure probability assumption, if \mathcal{D} is infinite then $H_P(\mathcal{D})$ coincides with the limit of the entropy of the truncated dictionary, completed with the failure set. For the latter entropy, we use the notation $H_P(\mathcal{D}_N)$, omitting the union with the failure set (just as in $L_P(\mathcal{D}_N)$).

Given $\delta > 0$, let

$$S_n^{(\delta)} \triangleq \{x^n \in \mathcal{A}^n \mid \exists a \in \mathcal{A}, s \in \mathcal{A}^k \text{ s.t. } n_s^{(a)}(x^n) < \delta n\}. \quad (43)$$

Thus, by our assumption that all conditional probabilities in processes $P \in \mathcal{P}_k$ are nonzero, for sufficiently small δ (depending on the specific P), $S_n^{(\delta)}$ is a large deviations event and thus its probability vanishes exponentially fast with n .

Our results are based on the following key lemma on dictionary entropies, where we use bounding techniques that are rooted in the source coding literature, particularly [31] and [32]. In the sequel, the $O(\cdot)$ notation refers to asymptotics relative to M . Thus, $O(1)$ denotes a quantity whose absolute value is upper-bounded by a constant, independent of M .

Lemma 9: Let $P \in \mathcal{P}_k$ and let \mathcal{D} be a dictionary.

- (i) If for every $x^* \in \mathcal{D}$ we have $|T(x^*)| \geq M$, then

$$H_P(\mathcal{D}_N) \geq \log M + (K/2) \log \log M - O(1) \quad (44)$$

for every $N > (\log M)/(\bar{H}_P(X) - \epsilon)$, where $\epsilon > 0$ is arbitrary. In addition, (44) also holds in the limit for $H_P(\mathcal{D})$.

- (ii) If for every $x^* \in \mathcal{D}$ we have $|T(x^*)| < CM$ for some positive constant C , except for a subset \mathcal{D}_0 of \mathcal{D} such that

$$\sum_{x^* \in \mathcal{D}_0} |x^*| P(x^*) = O(1), \quad (45)$$

then

$$H_P(\mathcal{D}_N) \leq \log M + (K/2) \log \log M + O(1) \quad (46)$$

for every $N > 0$. In addition, (46) also holds in the limit for $H_P(\mathcal{D})$.

Proof: Consider the universal sequential probability assignment $Q(x^*)$ on \mathcal{A}^* given by a uniform mixture over \mathcal{P}_k (namely, Laplace's rule of succession applied state by state), for which it is readily verified that

$$Q(x^n) = \prod_{s \in \mathcal{A}^k} \frac{\prod_{a \in \mathcal{A}} n_s^{(a)}(x^n)!}{(n_s(x^n) + \alpha - 1)!} (\alpha - 1)!.$$

It follows from Whittle's formula [17] that

$$Q(x^n) = \frac{W(x^n)}{|T(x^n)|} \prod_{s \in \mathcal{A}^k} \binom{n_s(x^n) + \alpha - 1}{\alpha - 1}^{-1} \quad (47)$$

where $W(x^n)$ denotes a determinant in the formula ("Whittle's cofactor") that satisfies $0 < W(x^n) \leq 1$, and accounts for certain restrictions in the state transitions which limit the universe of possible sequences that have the same type as x^n . We use this probability assignment to write the entropy of a generic, finite dictionary \mathcal{D}' , as

$$H_P(\mathcal{D}') = -\mathbf{E}_{P, \mathcal{D}'} \left[\log \frac{P(X^*)}{Q(X^*)} \right] + \mathbf{E}_{P, \mathcal{D}'} \left[\log \frac{1}{Q(X^*)} \right], \quad (48)$$

¹²The situation is akin to lossless source coding, for which the entropy bound is shown to be achievable with, say, the Shannon code, rather than with the (optimal) Huffman code.

where the first summation is the divergence between P and Q as distributions over \mathcal{D}' , which we denote $D_{\mathcal{D}'}(P||Q)$. By (47), we obtain

$$H_P(\mathcal{D}') = -D_{\mathcal{D}'}(P||Q) + \mathbf{E}_{P,\mathcal{D}'} \left[\log |T(X^*)| - \log W(X^*) + \sum_{s \in \mathcal{A}^k} \log \binom{n_s(X^*) + \alpha - 1}{\alpha - 1} \right]. \quad (49)$$

To prove Part (i), we first notice that if \mathcal{D}'' is also a dictionary and $\mathbf{T}_{\mathcal{D}''} \subseteq \mathbf{T}_{\mathcal{D}'}$, then it is easy to see that $H_P(\mathcal{D}'') \leq H_P(\mathcal{D}')$. Therefore, it suffices to prove the lemma for $N = N_1$, where

$$N_1 \triangleq \left\lceil \frac{\log M}{\bar{H}_P(X) - \epsilon} \right\rceil. \quad (50)$$

For convenience, the dictionary $\mathcal{D}_{N_1 \cup \mathcal{E}_{N_1}}$ is denoted \mathcal{D}' . Now, if $x^n \in \mathcal{D}'$, then either $|T(x^n)| \geq M$ or $n = N_1$, implying that

$$\begin{aligned} \mathbf{E}_{P,\mathcal{D}'} \log |T(X^*)| &\geq \left[1 - P \left(|T(x^{N_1})| < 2^{N_1(\bar{H}_P(X) - \epsilon)} \right) \right] \log M. \end{aligned} \quad (51)$$

Since the probability on the right-hand side of (51) decays exponentially fast with N_1 (see proof of Theorem 7), we obtain

$$\mathbf{E}_{P,\mathcal{D}'} \log |T(X^*)| \geq \log M - O((\log M)/M). \quad (52)$$

Next, recalling the definition of $N_0(M)$ in (29), for every $x^* \in \mathcal{D}'$ and $s \in \mathcal{A}^k$, $x^{N_0(M)}$ is a prefix of x^* , and therefore $n_s(x^*) \geq n_s(x^{N_0(M)})$. If $x^{N_0(M)} \notin S_{N_0(M)}^{(\delta)}$ (recall (43)), we have $n_s(x^{N_0(M)}) > \alpha \delta N_0(M) \geq (\alpha \delta \log M)/(\log \alpha)$. Thus, by Stirling's approximation, sequences $x^{N_0(M)} \in \mathcal{A}^{N_0(M)} \setminus S_{N_0(M)}^{(\delta)}$ satisfy

$$\sum_{s \in \mathcal{A}^k} \log \binom{n_s(x^{N_0(M)}) + \alpha - 1}{\alpha - 1} > K \log \log M - O(1). \quad (53)$$

Since $N_0(M) = \Omega(\log M)$ we have, for sufficiently small δ , $P(S_{N_0(M)}^{(\delta)}) = O(1/M)$. Thus, the right-hand side of (53) is also a lower bound on $\mathbf{E}_{P,\mathcal{D}'} \sum_{s \in \mathcal{A}^k} \log \binom{n_s(X^*) + \alpha - 1}{\alpha - 1}$. Finally, we have $\log W(x^*) \leq 0$ for every $x^* \in \mathcal{D}'$, so we conclude from (49), (52), and (53), that

$$H_P(\mathcal{D}') \geq -D_{\mathcal{D}'}(P||Q) + \log M + K \log \log M - O(1), \quad (54)$$

where the $O(1)$ term depends on P .

As for the divergence term in (54), it is easy to see that, since $\mathbf{T}_{\mathcal{D}'}$ is a sub-tree of the balanced tree of depth N_1 , we have

$$D_{\mathcal{D}'}(P||Q) \leq D_{\mathcal{A}^{N_1}}(P||Q). \quad (55)$$

Applying the divergence estimate in [33] (which extends to Markov sources the results in [34] on the asymptotics of the redundancy of Bayes rules with continuous prior densities) to sequences of length N_1 , we conclude that the divergence term in (54) is upper-bounded by $(K/2) \log \log M + O(1)$, proving

(44).¹³ The claim on $H_P(\mathcal{D})$ follows from its definition as the limit of a nondecreasing sequence.

To prove Part (ii) we first notice that, since $H_P(\mathcal{D}_N)$ is nondecreasing in N , it suffices to prove the upper bound for sufficiently large N . Moreover, since extending dictionary strings by a finite amount cannot lower the entropy, it suffices to prove it for dictionaries \mathcal{D} such that the length of the shortest sequence in the dictionary is at least $N_2 \triangleq c \log M$ for some positive constant c . Therefore, proceeding as in (55) we have, for sufficiently large N ,

$$D_{\mathcal{D}_N \cup \mathcal{E}_N}(P||Q) \geq D_{\mathcal{A}^{N_2}}(P||Q) = \frac{K}{2} \log \log M + O(1), \quad (56)$$

where the estimate follows, again, from [33, Corollary 1].

Next, we observe that, for $x^n \in \mathcal{D} \setminus \mathcal{D}_0$, we have $|T(x^n)| < CM$, whereas for $x^n \in \mathcal{D}_0$ we can use the trivial bound $\log |T(x^n)| < n \log \alpha$. Therefore, by (45),

$$\mathbf{E}_{P,\mathcal{D}} \log |T(X^*)| < \log M + O(1). \quad (57)$$

Similarly, for $n > N_1(C) \triangleq (\log(CM))/(\bar{H}_P(X) - \epsilon)$ for some $\epsilon > 0$, $\{x^n \in \mathcal{D} \setminus \mathcal{D}_0\}$ is a large deviations event and its probability decreases exponentially fast with n , as the type class size for a typical sequence will be at least CM . Thus, using again Stirling's approximation and (45), we obtain

$$\begin{aligned} \sum_{s \in \mathcal{A}^k} \mathbf{E}_{P,\mathcal{D}} \log \binom{n_s(X^*) + \alpha - 1}{\alpha - 1} &< K \log N_1(C) + O(1) \\ &= K \log \log M + O(1). \end{aligned} \quad (58)$$

Clearly, since $P(\mathcal{E}_N)$ vanishes as N grows, the upper bounds (57) and (58) hold, *a fortiori*, when the expectations are taken over sequences in $\mathcal{D}_N \cup \mathcal{E}_N$ instead of \mathcal{D} , for any $N > 0$.

Finally, for sequences $x^n \in \mathcal{A}^n \setminus S_n^{(\delta)}$ for some $\delta > 0$, $W(x^n)$ is known to be lower-bounded by a positive function of δ (see, e.g., [26, proof of Lemma 3]). For sequences $x^n \in S_n^{(\delta)}$ (an event whose probability decreases with n exponentially fast for sufficiently small δ), $W(x^n)$ is $\Omega(1/n^k)$ [35]. Therefore, $\mathbf{E}_{P,\mathcal{D}_N \cup \mathcal{E}_N} \log(1/W(X^*)) = O(1)$ for any $N > 0$. The upper bound (46) then follows from (49), (56), (57), and (58), both for $H_P(\mathcal{D}_N)$ and for $H_P(\mathcal{D})$. ■

To apply Lemma 9 to the estimation of the expected dictionary length, we need to link $L_P(\mathcal{D}_N)$ to $H_P(\mathcal{D}_N)$. Applying the LANSIT (recall (24)) to the self-information function, we obtain the “leaf-entropy theorem” (see, e.g., [36]), which states, for a generic dictionary \mathcal{D} , that

$$H_P(\mathcal{D}) = \sum_{x^* \in \mathbf{I}_{\mathcal{D}}} P(x^*) H_P(X|s(s_0, x^*)),$$

where $s(s_0, x^*)$ denotes the state assumed by the source after emitting x^* , starting at s_0 . In the memoryless case, $H_P(X|s(s_0, x^*))$ is independent of x^* , and further applying the LANSIT to the length function (as in (27)), we obtain $H_P(\mathcal{D}) = \bar{H}_P(X) L_P(\mathcal{D})$. This relation directly provides the desired link, and is used, e.g., in [8]. The situation is

¹³While the claim in [33, Corollary 1] is for a source in stationary mode, its proof actually builds on showing the same result for a fixed initial state, as in our setting.

more intricate for sources with memory, for which, regrouping terms, the theorem clearly takes the form

$$H_P(\mathcal{D}) = \sum_{t \in \mathcal{A}^k} H_P(X|t) \sum_{x^* \in \mathcal{I}_{\mathcal{D}}} P(x^*) \delta(t, s(s_0, x^*)),$$

where $\delta(t, s(s_0, x^*)) = 1$ if $t = s(s_0, x^*)$, and 0 otherwise. An additional application of the LANSIT, this time to the function $n_t(x^*)$, then yields

$$H_P(\mathcal{D}) = \sum_{t \in \mathcal{A}^k} H_P(X|t) \mathbf{E}_{P, \mathcal{D}} n_t(X^*). \quad (59)$$

A variant of this problem is studied in [31] in the context of variable-to-fixed-length source coding. We will make use of a result in [31], for which we need to consider the *extended* source defined on the strings of \mathcal{D} (referred to as a “segment source” in [31]), which is clearly also Markov with the same state set \mathcal{A}^k . Let $P^{\text{seg}}(s)$ denote its steady-state distribution when the chain is started with the stationary distribution of the basic (non-extended) source.¹⁴ Letting $L_P^{\text{seg}}(\mathcal{D})$ denote the expected dictionary length when the distribution on the initial state is $P^{\text{seg}}(s)$, [31, Lemma 1] states that, for any state $t \in \mathcal{A}^k$,

$$\sum_{s \in \mathcal{A}^k} P^{\text{seg}}(s) \mathbf{E}_{P, \mathcal{D}} n_t(s, X^*) = P^{\text{st}}(t) L_P^{\text{seg}}(\mathcal{D}), \quad (60)$$

where $n_t(s, X^*)$ is the same as $n_t(X^*)$ but with the source starting at state s , rather than s_0 .¹⁵ Hence,

$$\begin{aligned} & \sum_{s \in \mathcal{A}^k} P^{\text{seg}}(s) H_{P, s}(\mathcal{D}) \\ &= \sum_{t \in \mathcal{A}^k} H_P(X|t) \sum_{s \in \mathcal{A}^k} P^{\text{seg}}(s) \mathbf{E}_{P, \mathcal{D}} n_t(s, X^*) \\ &= \sum_{t \in \mathcal{A}^k} H_P(X|t) P^{\text{st}}(t) L_P^{\text{seg}}(\mathcal{D}) = \bar{H}_P(X) L_P^{\text{seg}}(\mathcal{D}), \end{aligned} \quad (61)$$

where the first equality follows from (59), the second one from (60), and the third one from (2). Notice, however, that the expected dictionary length may, in general, be quite sensitive to the initial state. Therefore, it is not clear whether the rightmost side of (61), which involves $L_P^{\text{seg}}(\mathcal{D})$, can provide information on $L_P(\mathcal{D})$. In addition, a dictionary that satisfies the conditions of Lemma 9 for a given initial state s_0 may not satisfy these conditions for a different initial state. While this issue is less serious (as the class type size is not very sensitive to the initial state), a direct application of the bounds shown in the lemma to the left-most side of (61) is problematic.

Next, we present an auxiliary lemma that will address these problems. To state the lemma, we need to introduce some additional objects. Given two states $s, t \in \mathcal{A}^k$, consider the

¹⁴As noted in [31], the segment source may not be irreducible, and therefore $P^{\text{seg}}(s)$ is one of possibly multiple stationary distributions. Note also that, in general, $P^{\text{seg}}(s)$ need not coincide with $P^{\text{st}}(s)$, unless \mathcal{D} is a balanced tree.

¹⁵While the statement of [31, Lemma 1] uses the above specific stationary distribution, the proof in fact holds for any stationary distribution. It is essentially based on counting the number $N_j(t)$ of occurrences of t in a sequence composed of j source segments, for large j . Since the state distribution at the segment starting points converges to $P^{\text{seg}}(s)$, with probability one, $N_j(t) = j P^{\text{st}}(t) L_P^{\text{seg}}(\mathcal{D})$. On the other hand, for segments starting at state s , t occurs $\mathbf{E}_{P, \mathcal{D}} n_t(s, X^*)$ times in the limit.

set $\mathcal{D}_{t,s}$ of all sequences x^* such that $s = s(t, x^*)$ and no proper prefix of x^* has this property. Since $\mathcal{D}_{t,s}$ has the prefix property and since every state is reachable from any other state, it is a (full, infinite) dictionary with a failure probability that vanishes as the truncation level grows for any $P \in \mathcal{P}_k$. The expected dictionary length for $\mathcal{D}_{t,s}$ (where the probabilities are computed with an assumed initial state t), which is the mean first passage time from t to s , is finite (since all states are positive recurrent); we denote it by $\mathcal{L}_{t,s}$.

For sequence sets U and V , let $U \cdot V$ denote the set $\{uv \mid u \in U, v \in V\}$. If U and V are dictionaries, then $U \cdot V$ is a dictionary whose corresponding tree is obtained by “hanging” the tree corresponding to V from each of the leaves of the tree corresponding to U .

Lemma 10: Let $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$ be a collection of sets of dictionaries, one set for each state in \mathcal{A}^k , which satisfies the following property: For every pair of states s, t , if $\mathbb{D}_s \in \mathbb{D}_s$ then $\mathcal{D}_{t,s} \cdot \mathbb{D}_s \in \mathbb{D}_t$. Let L_s^* denote the infimum over \mathbb{D}_s of the expected dictionary length, where the expectation assumes the initial state s , $s \in \mathcal{A}^k$, and $P \in \mathcal{P}_k$ is arbitrary. Assume L_s^* is finite and let $\mathcal{D}_s^* \in \mathbb{D}_s$, $s \in \mathcal{A}^k$, denote a dictionary that attains this infimum within some $\epsilon > 0$. Then, for any $s \in \mathcal{A}^k$, we have

$$\min_{t \in \mathcal{A}^k} H_{P, t}(\mathcal{D}_t^*) + K_1 \leq \bar{H}_P(X) L_s^* \leq \max_{t \in \mathcal{A}^k} H_{P, t}(\mathcal{D}_t^*) + K_2$$

for some constants K_1 and K_2 that are independent of $\{\mathbb{D}_s\}$.

The proof of Lemma 10 is presented in Appendix E.

G. Performance: Tight Bounds

In view of Lemma 5 and (44) in Part (i) of Lemma 9, to obtain a lower bound on $L_P(\mathcal{D})$ for universal VFRs, it suffices to apply Lemma 10 to a suitable collection of sets $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$; the lower bound will translate to truncated dictionaries by typicality arguments similar to those employed in the proof of Lemma 9. Our lower bound is stated in Theorem 8 below.

Theorem 8: Let $\mathcal{V} = (\mathcal{D}, \Phi, M)$ be a universal VFR such that $P(\mathcal{E}_N) \xrightarrow{N \rightarrow \infty} 0$ for every $P \in \mathcal{P}_k$. Then, for every $P \in \mathcal{P}_k$ and every $N > (\log M)/(\bar{H}_P(X) - \epsilon)$, where $\epsilon > 0$ is arbitrary, we have

$$L_P(\mathcal{D}_N) \geq \frac{\log M + (K/2) \log \log M - O(1)}{\bar{H}_P(X)}. \quad (62)$$

Proof: By Lemma 5, $|T(x^*)| \geq M$ for every $x^* \in \mathcal{D}$. Let \mathcal{D}' denote the dictionary obtained by “pruning” $\mathcal{T}_{\mathcal{D}}$ as follows: Replace every $x^* \in \mathcal{D}$ such that $|x^*| > N_1$ (where N_1 is given in (50) with ϵ an arbitrary positive constant) by its shortest prefix of length at least N_1 whose type class contains at least M elements. We first prove that the expected length of \mathcal{D}' satisfies the claimed lower bound.

To this end, consider the collection of dictionary sets $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$, where \mathbb{D}_s is the set of dictionaries \mathcal{D}_s such that, for an initial state s , $|T(x^*)| \geq M$ for every $x^* \in \mathcal{D}_s$ and $P(\mathcal{E}_N) \xrightarrow{N \rightarrow \infty} 0$ for every $P \in \mathcal{P}_k$. We show that this collection satisfies the conditions of Lemma 10. It is easy to see that L_s^* (where we use the notation introduced in Lemma 10) is finite for all $s \in \mathcal{A}^k$ (in fact, by typicality arguments, $L_s^* \leq N_1$) although, in any case, the claimed lower bound on $L_P(\mathcal{D}')$

would be trivial if L_s^* were infinite. To see that if $\mathcal{D}_s \in \mathbb{D}_s$ then $\mathcal{D}_{t,s} \cdot \mathcal{D}_s \in \mathbb{D}_t$ for every pair of states s, t , it suffices to observe that for a sequence $x^* = uv$ such that u takes the source from state t to state s , if $v' \in T(v)$ (where the type class assumes an initial state s) then $uv' \in T(x^*)$ (where the type class assumes an initial state t). Therefore, if $v \in \mathcal{D}_s$, we conclude that $|T(x^*)| \geq M$ for $x^* \in \mathcal{D}_{t,s} \cdot \mathcal{D}_s$. Since, clearly, $\mathcal{D}_{t,s} \cdot \mathcal{D}_s$ also has vanishing failure probability, the collection $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$ indeed satisfies the conditions of Lemma 10.

Now, applying Part (i) of Lemma 9 to \mathcal{D}_t^* , we obtain

$$H_{P,t}(\mathcal{D}_t^*) \geq \log M + (K/2) \log \log M - O(1)$$

for every $t \in \mathcal{A}^k$. It then follows from Lemma 10 that

$$L_{s_0}^* \geq \frac{\log M + (K/2) \log \log M - O(1)}{\bar{H}_P(X)}.$$

Since $\mathcal{D}' \in \mathbb{D}_{s_0}$ and $L_{s_0}^*$ is the infimum over \mathbb{D}_{s_0} of the expected dictionary length, the proof of our claim on \mathcal{D}' is complete.

Next, we observe that since for every internal node x^* at depth larger than N_1 of the tree corresponding to \mathcal{D}' we have $|T(x^*)| < M$, then

$$L_P(\mathcal{D}') - L_P(\mathcal{D}_{N_1}) \leq \sum_{i=0}^{\infty} (i+1) P(|T(x^{N_1+i})| < M).$$

Therefore, our claim on $L_P(\mathcal{D}_N)$ follows from our lower bound on $L_P(\mathcal{D}')$ by the typicality arguments used in the proof of Lemma 9, by which the probability that the type class of a sequence be smaller than $2^{N_1(\bar{H}_P(X) - \epsilon)}$ decays exponentially fast with the sequence length $N_1 + i$. ■

Remark 12: Although Theorem 8 is stated for a universal VFR, it is easy to see that it applies, like Lemma 5, also to perfect VFRs for “almost all” $P \in \mathcal{P}_k$. At first sight, this fact may seem surprising, since perfect VFRs for arbitrary memoryless distributions $P \in \mathcal{P}_0$ with expected dictionary length of the form $(\log M + O(1))/\bar{H}_P(X)$ are described in [7] and [8], where it is also shown that these VFRs are optimal up to a constant term. However, notice that, unlike the setting in Theorem 8, these VFRs are not required to be perfect at all truncation levels. Thus, in the memoryless case, the extra cost incurred by requiring perfection at all truncation levels is at least $(\log \log M)/(2\bar{H}_P(X))$. Since, as will be shown in the sequel, the lower bound (62) is achievable, the above minimum value of the extra cost is also achievable. The situation in the case $k > 0$, for which [8] proposes a VFR but does not provide tight bounds, is discussed later.

Next, we show that the lower bound (62) is achievable. To this end, we use Part (ii) of Lemma 9, for which we need a universal VFR such that the size of the type class of each sequence in its dictionary is at most CM for some constant C , except for a negligible subset of dictionary members. Such a bound on the type class size does not appear to follow easily from the definition of the optimal VFR \mathcal{V}^* since, in principle, the construction may require $|T| > CM$ for any constant C to guarantee $|\mathcal{D}^*(T)| \geq M$. Therefore, we will use a different universal VFR to show achievability.

To describe this universal VFR we will make use of an auxiliary dictionary $\tilde{\mathcal{D}}$, given by

$$\tilde{\mathcal{D}} = \{x^* \mid |S_u^-(T(x^*))| \geq M \ \forall u \in \mathcal{A}^{k+1}, \text{ and no } x^{**} \prec x^* \text{ has this property}\}, \quad (63)$$

where \prec denotes the proper prefix relation. Thus, $\tilde{\mathcal{D}}$ grows until the first time *each* set $S_u^-(T(x^*))$ (which, by Lemma 7 and Remark 11, if not empty, are distinct type classes in \mathcal{T}_{n-k-1}) is large enough.

Remark 13: The “stopping set” S defined in [9, Section IV] is the special case of $\tilde{\mathcal{D}}$ for the class of Bernoulli sources.

We first show that Part (ii) of Lemma 9 is applicable to $\tilde{\mathcal{D}}$.

Lemma 11: For every $P \in \mathcal{P}_k$, there exists a constant C such that $|T(x^*)| < CM$ for every $x^* \in \tilde{\mathcal{D}} \setminus \tilde{\mathcal{D}}_0$, where $\tilde{\mathcal{D}}_0$ is a subset of $\tilde{\mathcal{D}}$ satisfying

$$\sum_{x^* \in \tilde{\mathcal{D}}_0} |x^*| P(x^*) = O(1). \quad (64)$$

In addition, the probability of the failure set of $\tilde{\mathcal{D}}$ vanishes (exponentially fast).

Proof: Recalling the definition (43), let

$$\tilde{\mathcal{D}}_0 \triangleq \tilde{\mathcal{D}} \cap \left(\bigcup_{n \geq 1} S_n^{(\delta)} \right)$$

where $\delta > 0$ is small enough for $\{x^n \in S_n^{(\delta)}\}$ to be, by our assumptions on \mathcal{P}_k , a large deviations event, so its probability vanishes with n exponentially fast. Therefore, (64) holds.

Next, observe that, by Whittle’s formula for $|T(x^n)|$, if $x^n \notin S_n^{(\delta)}$ then deleting its last symbol can only affect the type class size by a multiplicative constant independent of n . As a result, for all u , $|T(x^n)| < \beta |S_u^-(T(x^n))|$ for some constant β (that depends on δ and $|u|$). Since, by (63) and (39), if $x^* \in \tilde{\mathcal{D}}$ then $|S_u^-(T(x^*))| < M$ for some $u \in \mathcal{A}^{k+2}$, we conclude that for every sequence $x^* \in \tilde{\mathcal{D}} \setminus \tilde{\mathcal{D}}_0$ we have $|T(x^*)| < CM$ for some constant C , as claimed.

Finally, notice that the failure set of $\tilde{\mathcal{D}}$ at truncation level N consists of sequences x^N such that $|S_u^-(T(x^N))| < M$ for some $u \in \mathcal{A}^{k+1}$. The event $\{x^N \in S_N^{(\delta)}\}$ has exponentially vanishing probability. If $x^N \notin S_N^{(\delta)}$ then, as discussed, $|T(x^N)| < \beta |S_u^-(T(x^N))|$. Hence, clearly, $P(T(x^N)) < \beta P(S_u^-(T(x^N)))$. If $N > N_1$ (as defined in (50)), $P(|S_u^-(T(x^N))| < M)$ vanishes exponentially fast. Therefore, so does the probability of the failure set. ■

We can now use the upper bound (46) and Lemma 10 to upper-bound $L_P(\tilde{\mathcal{D}})$ as follows.

Lemma 12: For every $P \in \mathcal{P}_k$ we have

$$L_P(\tilde{\mathcal{D}}) \leq \frac{\log M + (K/2) \log \log M + O(1)}{\bar{H}_P(X)}. \quad (65)$$

Proof: Consider the collection $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$ where \mathbb{D}_s is the set of dictionaries \mathcal{D}_s such that, for an initial state s , if $x^* \in \mathcal{D}_s$ then $|S_u^-(T(x^*))| \geq M$ for all $u \in \mathcal{A}^{k+1}$. Clearly, the same arguments as in the proof of Theorem 8 prove that the collection $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$ satisfies the conditions of Lemma 10. By (63), the dictionary with shortest expected length over

\mathbb{D}_s is precisely $\tilde{\mathcal{D}}$ with initial state s , which we denote $\tilde{\mathcal{D}}_s$. Therefore, by Lemma 10,

$$\bar{H}_P(X)L_P(\tilde{\mathcal{D}}_s) \leq \max_{t \in \mathcal{A}^k} H_{P,t}(\tilde{\mathcal{D}}_t) + K_2$$

for some constant K_2 . Now, by Lemma 11 and Part (ii) of Lemma 9, we have

$$H_{P,t}(\tilde{\mathcal{D}}_t) \leq \log M + (K/2) \log \log M + O(1)$$

independently of the initial state, which completes the proof. \blacksquare

Thus, the expected length of $\tilde{\mathcal{D}}$ coincides (up to an additive constant) with the lower bound of Theorem 8 on the expected length of a universal VFR. Recall that Lemma 5 requires that, for the dictionary \mathcal{D} of a universal VFR, M divide $|\mathcal{D}(T(x^*))|$ for all $x^* \in \mathcal{D}$. While $\tilde{\mathcal{D}}$ may not satisfy this property, the following result makes it a suitable “building block” in the construction of a universal VFR.

Lemma 13: For every $x^* \in \tilde{\mathcal{D}}$ we have $|\tilde{\mathcal{D}}(T(x^*))| \geq M$.

Proof: Let $x^n \in \tilde{\mathcal{D}}$, $T \triangleq T(x^n)$, and $T' \triangleq T(x^{n-1})$. For a generic string $w \in \mathcal{A}^*$, consider the following properties:

- (P1) $|S_w^-(T')| \geq M$;
- (P2) For every suffix t of w , there exists $u \in \mathcal{A}^{k+1}$ such that $|S_{ut}^-(T')| < M$;
- (P3) $|S_{aw}^-(T')| < M$ for all $a \in \mathcal{A}$.

We claim the existence of a string w satisfying (P1)–(P3). We will exhibit such a string, by constructing a sequence $v^{(0)}, v^{(1)}, \dots, v^{(i)}, \dots$ of strings of strictly increasing length, each satisfying (P1)–(P2), and with the property that, given $v^{(i)}$, $i \geq 0$, there exists a string z such that either $w = zv^{(i)}$ satisfies (P1)–(P3) (and our claim is proven) or, for some $c \in \mathcal{A}$, $v^{(i+1)} = czv^{(i)}$ satisfies (P1)–(P2), and we can extend the sequence by one element. Such a construction cannot continue indefinitely without finding the desired string w , since, as the length of $v^{(i)}$ increases, eventually we would have $S_{v^{(i)}}^-(T') = \emptyset$, so $v^{(i)}$ would not satisfy (P1).

To construct the sequence $\{v^{(i)}\}$, we first establish that $v^{(0)} = \lambda$ satisfies (P1)–(P2). Since $T' = S_{x^{n-k}}^-(T)$ and $x^n \in \tilde{\mathcal{D}}$, by (63), $|T'| \geq M$, so (P1) holds for $w = \lambda$. But since $x^{n-1} \notin \tilde{\mathcal{D}}$, again by (63), (P2) also holds for λ . Next, to prove the existence of the mentioned string z given $v^{(i)}$, $i \geq 0$, we need the following lemma, which is proved in Appendix F.

Lemma 14: Let $T_1 \in \mathcal{T}_{n+1}$, $b, c \in \mathcal{A}$, $T_2 = S_b^-(T_1)$, and $T_3 = S_c^-(T_1)$. Then, there exists $u \in \mathcal{A}^{k+1}$ such that $|T_2| \geq |S_u^-(T_3)|$.

Assume that we are given $v^{(i)}$, $i \geq 0$, satisfying (P1)–(P2). By (P2), there exists $z' \in \mathcal{A}^{k+1}$ such that $|S_{z'v^{(i)}}^-(T')| < M$. *A fortiori*, every suffix of $z'v^{(i)}$ also satisfies (P2). Let z'' denote the shortest suffix of z' such that $|S_{z''v^{(i)}}^-(T')| < M$. Since $v^{(i)}$ satisfies (P1), we have $|z''| > 0$, so let $z'' = bz$, $b \in \mathcal{A}$. Now, if for every $c \in \mathcal{A}$ we have $|S_{czv^{(i)}}^-(T')| < M$, then (P1)–(P3) hold for $w = zv^{(i)}$, as claimed. Otherwise, let c be such that $|S_{czv^{(i)}}^-(T')| \geq M$. By Lemma 14 (with $T_1 = S_{zv^{(i)}}^-(T')$), there exists $r \in \mathcal{A}^{k+1}$ such that $|S_{rczv^{(i)}}^-(T')| \leq |S_{bzc}^-(T')| < M$. Hence, (P1)–(P2) hold for $v^{(i+1)} = rczv^{(i)}$. We have thus shown the existence of a string w satisfying (P1)–(P3).

-
- 1) Set $i = 0$, $\tilde{\mathcal{D}}^* = \emptyset$, and let $\Delta_0 = \tilde{\mathcal{D}}$. For $x^* \in \tilde{\mathcal{D}}$, define $\mathcal{G}_0(x^*) = \tilde{\mathcal{D}}(T(x^*))$.
 - 2) Set $\Delta_{i+1} = \emptyset$. For each $\mathcal{G} = \mathcal{G}_i(x^*)$, $x^* \in \Delta_i$, do:
 - a) Let $m = |\mathcal{G}| \bmod M$, and let U be a set of m sequences from \mathcal{G} . Add $\mathcal{G} \setminus U$ to $\tilde{\mathcal{D}}^*$.
 - b) If $m > 0$, let s_f be the common final state of all sequences in U . Add $U \cdot \tilde{\mathcal{D}}_{s_f}(\lceil M/m \rceil)$ to Δ_{i+1} .
 - 3) If $\Delta_{i+1} = \emptyset$, **stop**. Otherwise, for each $x^* \in \Delta_{i+1}$, let x^ℓ be its prefix in Δ_i , and define $\mathcal{G}_{i+1}(x^*) = \{y^* \in T(x^*) \cap \Delta_{i+1} \mid y^\ell \in \mathcal{G}_i(x^\ell)\}$. Increment i , and go to Step 2.
-

Fig. 5. Description of the universal VFR $\tilde{\mathcal{D}}^*$.

Next, let $y^n \in S_{v^{x_{n-k}}}(T)$, and denote $|w| = \ell$. By (39), $y^{n-\ell-1} \in S_v^-(T')$. Since w satisfies (P3), no prefix of $y^{n-\ell-1}$ satisfies the membership condition of (63). Consider now y^{n-j-1} , $0 \leq j < \ell$. Since $y^{n-1} \in S_w(T')$, it is easy to see that $y^{n-j-1} \in S_t^-(T')$, where t is the (proper) suffix of w of length j . Thus, by (P2), y^{n-j-1} does not satisfy the membership condition either. It follows that the membership condition is not satisfied for any proper prefix of y^n . But since $S_{wx_{n-k}}(T) \subseteq T$, the condition is satisfied for y^n and, hence, $y^n \in \tilde{\mathcal{D}}$. The proof is complete by noticing that $|S_{wx_{n-k}}(T)| = |S_w^-(T')|$ and invoking (P1) for w . \blacksquare

Next, we describe the construction of the dictionary $\tilde{\mathcal{D}}^*$ of a universal VFR. In this construction, the value of the initial state implicit in the class type definition in $\tilde{\mathcal{D}}$ will *not* be fixed at the same s_0 throughout, and the value of the threshold for the class type size used in (63) may differ from M . Therefore, it will be convenient to explicitly denote $\tilde{\mathcal{D}}$ by $\tilde{\mathcal{D}}_s(\ell)$, where s denotes the initial state and ℓ is the threshold, while maintaining the shorthand notation $\tilde{\mathcal{D}} \triangleq \tilde{\mathcal{D}}_{s_0}(M)$.

The dictionary $\tilde{\mathcal{D}}^*$ is iteratively described, as shown in Figure 5. At the beginning of the i th iteration, $\tilde{\mathcal{D}}^*$ contains a prefix set of sequences which have been added to the set in previous iterations. In addition, there is a prefix set Δ_i of sequences that are still pending processing (i.e., either inclusion in $\tilde{\mathcal{D}}^*$, or extension), such that $\tilde{\mathcal{D}}^* \cup \Delta_i$ is a full dictionary. Initially, $\tilde{\mathcal{D}}^*$ is empty and $\Delta_0 = \tilde{\mathcal{D}}$. Sequences in Δ_i are collected into groups $\mathcal{G}_i(x^*)$, where the latter consists of all the pending sequences of the same type as x^* , and whose prefixes in the previous iteration were in the same group $\mathcal{G}_{i-1}(x^*)$ (thus, recursively, the prefixes were also of the same type in all prior iterations).

The dictionary $\tilde{\mathcal{D}}^*$ is built up, in Step 2a, of sets of sizes divisible by M , consisting of sequences of the same type. Thus, M divides $|\tilde{\mathcal{D}}^*(T)|$ for all n and all type classes $T \in \mathcal{T}_n$, so that Lemma 5 guarantees the existence of a universal VFR based on $\tilde{\mathcal{D}}^*$. The remaining m sequences are recursively extended by “hanging,” in Step 2b, dictionaries $\tilde{\mathcal{D}}_{s_f}(\lceil M/m \rceil)$. Thus, by Lemma 13, the new set Δ_{i+1} contains m copies of type classes of sizes at least M/m . Unless, at some step i , $m = 0$ for all groups (i.e., Δ_{i+1} is empty), the procedure continues indefinitely and, as a result, $\tilde{\mathcal{D}}^*$ contains more infinite paths than $\tilde{\mathcal{D}}$. The choice of threshold $\lceil M/m \rceil$ in Step 2b guarantees the following property for the groups \mathcal{G}_i in Step 2a.

Lemma 15: For all $i \geq 0$ and all $x^* \in \Delta_i$, we have $|\mathcal{G}_i(x^*)| \geq M$.

Proof: Lemma 13 guarantees that the claim is true for $i = 0$ (Step 1 in Fig. 5). By Steps 1 and 3, sequences in the same group \mathcal{G}_i are indeed of the same type, and, thus, s_f is well defined in Step 2b, where Δ_{i+1} is built-up of subsets of the form $U \cdot \tilde{\mathcal{D}}_{s_f}(\lceil M/m \rceil)$. Also, by Lemma 13, the size of the type class of every sequence in $\tilde{\mathcal{D}}_{s_f}(\lceil M/m \rceil)$ is at least $\lceil M/m \rceil$. Now, sequences in the same group \mathcal{G}_i , ending in state s_f , when appended with sequences from $\tilde{\mathcal{D}}_{s_f}(\lceil M/m \rceil)$ that are of the same type with respect to the *initial* state s_f , remain in the same group \mathcal{G}_{i+1} (Step 2b and definition of \mathcal{G}_{i+1} in Step 3). In particular, this applies to the sequences in the set U , and, therefore, $|\mathcal{G}_{i+1}(x^*)| \geq m \lceil M/m \rceil \geq M$ for all $x^* \in \Delta_{i+1}$. ■

We now turn to the computation of $L_P(\tilde{\mathcal{D}}^*)$. We first show, in Lemma 16 below, that the iterative process described in Figure 5 does not increase the expected length of $\tilde{\mathcal{D}}$ by more than a constant.

Lemma 16: For every $P \in \mathcal{P}_k$ we have

$$L_P(\tilde{\mathcal{D}}^*) - L_P(\tilde{\mathcal{D}}) = O(1).$$

Proof: Let \mathbb{G}_i denote the partition of Δ_i into groups \mathcal{G}_i , and for each $\mathcal{G} \in \mathbb{G}_i$, let $m(\mathcal{G}) = |\mathcal{G}| \bmod M$. Let \mathcal{L}_i denote the limiting expected truncated length of the full, prefix set formed by the union of Δ_i and the “current state” of $\tilde{\mathcal{D}}^*$ after the i th iteration of the algorithm. Clearly,

$$L_P(\tilde{\mathcal{D}}^*) = \lim_{i \rightarrow \infty} \mathcal{L}_i. \quad (66)$$

Since all the sequences in a group \mathcal{G} are equiprobable and only $m(\mathcal{G})$ of them are extended in Step 2b whereas, by Lemma 15, at least M of them are not, we have

$$\mathcal{L}_{i+1} \leq \mathcal{L}_i + \sum_{\mathcal{G} \in \mathbb{G}_{i+1}} \frac{m(\mathcal{G})}{M + m(\mathcal{G})} P(\mathcal{G}) L_P(\tilde{\mathcal{D}}_{s_f}(\lceil M/m(\mathcal{G}) \rceil)) \quad (67)$$

where s_f denotes the common final state of the sequences in \mathcal{G} . By Lemma 12, applied rather loosely, for every $s \in \mathcal{A}^k$ and every positive integer ℓ , we have $L_P(\tilde{\mathcal{D}}_s(\ell)) < \eta \ell$ for some constant η , independent of M . Therefore, (67) implies

$$\mathcal{L}_{i+1} \leq \mathcal{L}_i + \eta \sum_{\mathcal{G} \in \mathbb{G}_{i+1}} P(\mathcal{G}) = \mathcal{L}_i + \eta P(\Delta_{i+1}). \quad (68)$$

Now, since $m(\mathcal{G}) < M$, it follows from Lemma 15 that more than half of the pending sequences make it to $\tilde{\mathcal{D}}^*$ in Step 2a. Thus, since Δ_i is a prefix set, we have $P(\Delta_i) > 2P(\Delta_{i+1})$ and, hence,

$$P(\Delta_{i+1}) < 2^{-(i+1)}.$$

It then follows from (68) that

$$\mathcal{L}_i < \mathcal{L}_0 + 2\eta = L_P(\tilde{\mathcal{D}}) + 2\eta$$

which, together with (66), completes the proof. ■

Therefore, just as $\tilde{\mathcal{D}}, \tilde{\mathcal{D}}^*$ attains the lower bound (62). The following characterization of $L_P(\mathcal{D}_N^*)$ then follows straightforwardly from Theorem 8, Lemma 12, Lemma 16, and Theorem 6.

Theorem 9: For every $P \in \mathcal{P}_k$ and sufficiently large N , the truncated dictionary \mathcal{D}_N^* of the optimal universal TVFR \mathcal{V}_N^* satisfies

$$L_P(\mathcal{D}_N^*) = \frac{\log M + (K/2) \log \log M + O(1)}{\bar{H}_P(X)}. \quad (69)$$

Remark 14: The term $(K/2) \log \log M$ in (69) resembles a typical “model cost” term in universal lossless compression but, as mentioned in Remark 12, the universality of the VFR does not entail an extra cost. Instead, in the memoryless case, $(\log \log M)/(2\bar{H}_P(X))$ is, as follows from the results in [7] and [8] discussed in Remark 12, the extra cost of maintaining perfection under truncation, in either the universal or individual process cases. The case $k > 0$ is also discussed in [8], but the bounds provided in that work are not tight enough to reach a similar conclusion. Nevertheless, Lemma 10 yields a tighter lower bound on the expected dictionary length for any perfect VFR without the truncation requirement. To derive this bound, let \mathbb{D}_s be the set of dictionaries \mathcal{D}_s corresponding to such VFRs for an initial state $s \in \mathcal{A}^k$. Clearly, $\{\mathbb{D}_s\}_{s \in \mathcal{A}^k}$ satisfies the conditions of Lemma 10. Since the dictionary strings can be clustered into M groups, each of probability $1/M$ (in the limit), we have $H_{P,s}(\mathcal{D}_s) \geq \log M$ for any $\mathcal{D}_s \in \mathbb{D}_s$ and any $s \in \mathcal{A}^k$. Thus, the lemma implies that $(\log M - O(1))/\bar{H}_P(X)$ is a lower bound on the expected dictionary length for any perfect VFR (without the truncation requirement). We conclude that, for $k > 0$, the extra cost of maintaining perfection under truncation, in either the universal or individual process cases, is *at most* $(K/2) \log \log M$. The question of whether this value is also a lower bound remains open, as it requires to improve on the upper bound provided in [8] on the expected dictionary length of the “interval algorithm.” The second order asymptotic analysis of the performance of universal VFRs on which no truncation requirements are posed also remains an open question.

APPENDIX A PROOF OF LEMMA 1

Before we proceed with the proof, we define the subset Ψ_0 mentioned in the statement of the lemma. Consider functions $g(\mathbf{p}) = \sum_{T \in \mathcal{T}_n} g_T P(T)/|T|$, where the g_T are integers, and $P(T)$ is the total probability of the type class T for a parameter $\mathbf{p} \in \Psi$. These functions are multivariate polynomials in the components of \mathbf{p} . Let

$$G_n = \{g(\mathbf{p}) \mid |g_T| \leq |T| \ \forall T \in \mathcal{T}_n, \ g_T \neq 0 \text{ for some } T \in \mathcal{T}_n\}. \quad (70)$$

It is known (see, e.g., [22], [23]) that the type probabilities $P(T)$, as functions of \mathbf{p} , are linearly independent over the reals. Thus, no $g \in G_n$ is identically zero. Let Ψ_0 denote the set of all vectors \mathbf{p} such that $g(\mathbf{p}) = 0$ for some $g \in G_n$. It is readily verified that Ψ_0 has volume zero in Ψ .

Proof of Lemma 1: Let $P_{\mathbf{p}}$ be a process in \mathcal{P}_k , where we use $P_{\mathbf{p}}$ instead of P to emphasize the dependence of the probabilities on the parameter vector $\mathbf{p} \in \Psi$. Consider a pair $(r, M) \in \mathbb{N}^+ \times \mathbb{N}_t$ such that $r \in [M]$ and $\gamma \triangleq P_{\mathbf{p}}(\mathcal{M}(x^n) = M) \neq 0$, and let χ denote the set of sequences x^n such that

$\mathcal{M}(x^n) = M$ and $\rho(x^n) = r$. Since sequences of the same type are equiprobable, we have

$$\begin{aligned} P_{\mathbf{p}}(\rho(X^n)=r \mid \mathcal{M}(X^n)=M) &= \gamma^{-1} \sum_{x^n \in \chi} P_{\mathbf{p}}(x^n) \\ &= \gamma^{-1} \sum_{T \in \mathcal{T}_n} \sum_{x^n \in \chi \cap T} P(x^n) = \gamma^{-1} \sum_{T \in \mathcal{T}_n} \frac{|\chi \cap T|}{|T|} P_{\mathbf{p}}(T). \end{aligned} \quad (71)$$

If the condition of the lemma holds, then the right-hand side of (71) is independent of r and, thus, \mathcal{F}_n is universal. Conversely, if \mathcal{F}_n is perfect for some $\mathbf{p} \in \Psi$, it follows from (71) that for any $r, r' \in [M]$, we have

$$\sum_{T \in \mathcal{T}_n} \frac{|\chi \cap T| - |\chi' \cap T|}{|T|} P_{\mathbf{p}}(T) = 0, \quad (72)$$

where χ' is defined as χ , but for r' . If the condition of the lemma does not hold, then, for some choice of $M, r, r' \in [M]$, and T , we have $|\chi \cap T| \neq |\chi' \cap T|$, and the expression on the left-hand side of (72), viewed as a multivariate polynomial in the components of \mathbf{p} , belongs to G_n . Thus, by the definition of Ψ_0 , we must have $\mathbf{p} \in \Psi_0$. ■

APPENDIX B PROOF OF LEMMA 2

Proof: We prove the lemma by induction on m . For $m = 1$, we have $H = 0$ and the claim is trivial. For $m > 1$, define $\mathbf{q}' = [q'_1, q'_2, \dots, q'_{m-1}]$ with $q'_i = q_{i+1}/(1 - q_1)$, $1 \leq i < m$. We claim that \mathbf{q}' satisfies the version of (10) for distributions on $m - 1$ symbols. Indeed, for $1 \leq i < m$, we have

$$c q'_i = \frac{c q_{i+1}}{1 - q_1} \geq \frac{1}{1 - q_1} \left(1 - \sum_{j=1}^i q_j\right) = 1 - \sum_{j=1}^{i-1} q'_j,$$

where the first inequality follows from the assumptions of the lemma, and the last equality from the definition of \mathbf{q}' and some algebraic manipulations. Now, denoting by H' the entropy of \mathbf{q}' , we can write

$$\begin{aligned} H &= h(q_1) + (1 - q_1)H' \leq h(q_1) + (1 - q_1)c h(c^{-1}) \\ &= q_1 \left(\frac{h(q_1)}{q_1} - \frac{h(c^{-1})}{c^{-1}} \right) + c h(c^{-1}) \leq c h(c^{-1}), \end{aligned}$$

where the first inequality follows from the induction hypothesis, and the second inequality from the fact that, by (10), we have $q_1 \geq c^{-1}$ and the function $h(x)/x$ is monotonically decreasing. ■

APPENDIX C PROOF OF LEMMA 4

Proof of Lemma 4: The proof is by induction on $|\mathcal{B}|$. The claim is trivial for $|\mathcal{B}| = 1$. Assume it also holds for all $|\mathcal{B}| < B$, where $B > 1$. Then, if $|\mathcal{B}| = B$, letting u_0 denote

an element with maximum probability in \mathcal{B} , we have

$$\begin{aligned} \sum_{u, v \in \mathcal{B}} |R(u) - R(v)| &= \sum_{u, v \in \mathcal{B} \setminus \{u_0\}} |R(u) - R(v)| + 2 \sum_{v \in \mathcal{B} \setminus \{u_0\}} (R(u_0) - R(v)) \\ &\leq 2(B - 2)(R(\mathcal{B}) - R(u_0)) + 2(BR(u_0) - R(\mathcal{B})) \\ &= 4R(u_0) + 2(B - 3)R(\mathcal{B}) \leq 2(B - 1)R(\mathcal{B}), \end{aligned}$$

where the first inequality follows from the induction hypothesis and the last one from $R(u_0) \leq R(\mathcal{B})$. ■

APPENDIX D PROOF OF LEMMA 5

Proof of Lemma 5: Consider the VFR $\mathcal{V} = (\mathcal{D}, \Phi, M)$. For a type class $T \in \mathcal{T}_n$, and $r \in [M]$, define the set $\mathcal{D}(T)_r = \{x^n \in \mathcal{D}(T) \mid \Phi(x^n) = r\}$. Assume that \mathcal{V} satisfies the condition of the lemma, i.e., that $|\mathcal{D}(T)_r|$ is independent of r for all n and all $T \in \mathcal{T}_n$. We claim that \mathcal{V}_N is universal at all truncation levels N (and, thus, \mathcal{V} is universal). Indeed, letting, for conciseness, $P_{\mathbf{p}}(r, N)$ denote the probability $P_{\mathbf{p}}(\Phi(X^*) = r, X^* \in \mathcal{D}_N)$, where, again, we emphasize the dependence of a process $P_{\mathbf{p}} \in \mathcal{P}_k$ on its parameter \mathbf{p} , we have

$$\begin{aligned} P_{\mathbf{p}}(r, N) &= \sum_{n=1}^N \sum_{\substack{x^n \in \mathcal{D}_N, \\ \Phi(x^n)=r}} P_{\mathbf{p}}(x^n) \\ &= \sum_{n=1}^N \sum_{T \in \mathcal{T}_n} \sum_{x^n \in \mathcal{D}(T)_r} P_{\mathbf{p}}(x^n) = \sum_{n=1}^N \sum_{T \in \mathcal{T}_n} \frac{|\mathcal{D}(T)_r|}{|T|} P_{\mathbf{p}}(T). \end{aligned} \quad (73)$$

The expression on the right-hand side of (73) is independent of r , establishing our claim. Assume now that \mathcal{V} does not satisfy the condition of the lemma, and let N be the smallest integer for which a type $T' \in \mathcal{T}_N$ violates the condition, i.e., for some $r, r' \in [M]$, we have $|\mathcal{D}_N(T')_r| \neq |\mathcal{D}_N(T')_{r'}|$. Consider a process $P_{\mathbf{p}}$ such that \mathcal{V}_N is perfect for $P_{\mathbf{p}}$. By (73), applied also to r' , we have

$$\begin{aligned} P_{\mathbf{p}}(r, N) - P_{\mathbf{p}}(r', N) &= \sum_{n=1}^N \sum_{T \in \mathcal{T}_n} \frac{|\mathcal{D}(T)_r| - |\mathcal{D}(T)_{r'}|}{|T|} P_{\mathbf{p}}(T) \\ &= \sum_{T \in \mathcal{T}_N} \frac{|\mathcal{D}(T)_r| - |\mathcal{D}(T)_{r'}|}{|T|} P_{\mathbf{p}}(T), \end{aligned} \quad (74)$$

where the last equality follows from our assumption on N . Also from the same assumption, it follows that at least one of the numerators in the expression on the rightmost side of (74) is nonzero, and the expression, as a multivariate polynomial in the entries of \mathbf{p} , belongs to G_N as defined in (70). We now reason as in the proof of Lemma 1 to conclude that \mathbf{p} must belong to a (fixed) subset Ψ'_0 of measure zero in Ψ . ■

APPENDIX E
PROOF OF LEMMA 10

Proof of Lemma 10: We will denote by $L_{P,s}$ the expected dictionary length under P , with initial state s . For every pair of states $t, s \in \mathcal{A}^k$, since $\mathcal{D}_{t,s} \cdot \mathcal{D}_s^* \in \mathbb{D}_t$, we have

$$L_t^* \leq L_{P,t}(\mathcal{D}_{t,s} \cdot \mathcal{D}_s^*) = \mathcal{L}_{t,s} + L_{P,s}(\mathcal{D}_s^*) \leq L_s^* + \epsilon + \mathcal{L}_{t,s}.$$

Therefore,

$$|L_s^* - L_t^*| \leq \max_{u,v \in \mathcal{A}^k} \mathcal{L}_{u,v} + \epsilon. \quad (75)$$

Now, let $\mathcal{D}_{s,N}^*$ denote the truncation of \mathcal{D}_s^* to depth N (completed with the corresponding failure set), and let $\mathcal{D}_N^{\text{univ}}$ denote the dictionary given by $\{su_s \mid s \in \mathcal{A}^k, u_s \in \mathcal{D}_{s,N}^*\}$. Thus, $\mathbf{T}_{\mathcal{D}_N^{\text{univ}}}$ is obtained by taking a balanced tree of depth k , and “hanging” from each leaf s the tree $\mathbf{T}_{\mathcal{D}_{s,N}^*}$, $s \in \mathcal{A}^k$. Clearly,

$$L_{P,s}(\mathcal{D}_N^{\text{univ}}) = k + \sum_{t \in \mathcal{A}^k} P^{(k)}(t|s) L_{P,t}(\mathcal{D}_{t,N}^*), \quad (76)$$

where $P^{(k)}(t|s)$ denotes the probability of moving from state s to state t in k steps. Since, for large enough N , $L_{P,t}(\mathcal{D}_{t,N}^*)$ is arbitrarily close to L_t^* , (75) and (76) imply that, for every $s, t \in \mathcal{A}^k$,

$$|L_{P,s}(\mathcal{D}_N^{\text{univ}}) - L_t^*| < C_1, \quad (77)$$

where the constant C_1 depends on P but is independent of s and t . Similarly,

$$H_{P,s}(\mathcal{D}_N^{\text{univ}}) = H_{P,s}(X^k) + \sum_{t \in \mathcal{A}^k} P^{(k)}(t|s) H_{P,t}(\mathcal{D}_{t,N}^*)$$

where $H_{P,s}(X^k)$ denotes the entropy of k -tuples (starting at state s). Therefore, applying (61) to $\mathcal{D}_N^{\text{univ}}$, we obtain

$$\begin{aligned} \bar{H}_P(X) L_P^{\text{seg}}(\mathcal{D}_N^{\text{univ}}) \\ = C_2 + \sum_{s,t \in \mathcal{A}^k} P^{\text{seg}}(s) P^{(k)}(t|s) H_{P,t}(\mathcal{D}_{t,N}^*) \end{aligned} \quad (78)$$

where C_2 is a constant that depends only on P . Together with (77), and taking the limit as $N \rightarrow \infty$ so that $H_{P,t}(\mathcal{D}_{t,N}^*) \rightarrow H_{P,t}(\mathcal{D}_t^*)$ (since the failure probability vanishes), (78) implies the claim of the lemma. ■

APPENDIX F
PROOF OF LEMMA 14

Proof of Lemma 14: Let z_1^k denote the final state of T_1 . We prove the lemma with $u = bz_1^k$. If $S_u^-(T_3)$ is empty, the lemma holds trivially. Otherwise, the sequences in T_1 include at least one transition from state bz_1^{k-1} to state z_1^k . It is easy to see that there exists a sequence in T_1 such that one of these transitions is the final one, and therefore T_2 is not empty. Thus, we assume that $T_2, T_3 \in \mathcal{T}_n$ and $S_u^-(T_3) \in \mathcal{T}_{n-k-1}$. The counts defining $S_u^-(T_3)$ differ from those defining T_1 in that a chain of state transitions

$$bz_1^{k-1} \rightarrow z_1^k \rightarrow z_2^k c \rightarrow z_3^k cz_1 \rightarrow \dots \rightarrow z_k cz_1^{k-2} \rightarrow cz_1^{k-1} \rightarrow z_1^k$$

has been deleted. On the other hand, T_2 is obtained by deleting from T_1 a transition $bz_1^{k-1} \rightarrow z_1^k$. Therefore, $S_u^-(T_3)$ can be obtained from T_2 by deleting a *circuit* of state transitions

$$z_1^k \rightarrow z_2^k c \rightarrow z_3^k cz_1 \rightarrow \dots \rightarrow z_k cz_1^{k-2} \rightarrow cz_1^{k-1} \rightarrow z_1^k. \quad (79)$$

At least one of the states, t_1^k , in the circuit (79), must occur in $x^{n-k-1} \in S_u^-(T_3)$, for otherwise the transition graph of T_2 , obtained by adding the circuit, would have a disconnected component. Fix $x^{n-k-1} \in S_u^-(T_3)$, and let i be such that $x_{i+1}^{i+k} = t_1^k$ is the last occurrence of this state in the sequence, $0 \leq i \leq n-2k-1$, and let v_1^{k+1} denote the string of symbols determined by the circuit (79) starting at a transition from state t_1^k (and ending at the same state). Clearly, the sequence $x_{i+k}^{i+k} v_1^{k+1} x_{i+k+1}^{n-k-1}$ is in T_2 , as the inserted string generates all the missing transitions prescribed by (79), returning to state t_1^k . In addition, it is easy to see that, with this procedure, two different sequences in $S_u^-(T_3)$ generate two different sequences in T_2 , which completes the proof. ■

REFERENCES

- [1] J. von Neumann, “Various techniques used in connection with random digits,” *U.S. Nat. Bur. Standards, Appl. Math. Series*, vol. 12, pp. 36–38, Jan. 1951.
- [2] P. Elias, “The efficient construction of an unbiased random sequence,” *Ann. Math. Statist.*, vol. 43, pp. 865–870, 1972.
- [3] Y. Peres, “Iterating von Neumann’s procedure for extracting random bits,” *Annals of Stat.*, vol. 20, no. 1, pp. 590–597, Mar. 1992.
- [4] S. Vembu and S. Verdú, “Generating random bits from an arbitrary source: fundamental limits,” *IEEE Trans. Inform. Theory*, vol. 41, no. 5, pp. 1322–1332, Sep. 1995.
- [5] W. Hoeffding and G. Simons, “Unbiased coin tossing with a biased coin,” *Ann. Math. Statist.*, vol. 41, no. 2, pp. 341–352, 1970.
- [6] Q. F. Stout and B. Warren, “Tree algorithms for unbiased coin tossing with a biased coin,” *Ann. Probab.*, vol. 12, no. 1, pp. 212–222, 1984.
- [7] J. R. Roche, “Efficient generation of random variables from biased coins,” in *Proc. 1991 International Symposium on Information Theory*, Budapest, Hungary, 1991.
- [8] T. S. Han and M. Hoshi, “Interval algorithm for random number generation,” *IEEE Trans. Inform. Theory*, vol. 43, no. 2, pp. 599–611, Mar. 1997.
- [9] H. Zhou and J. Bruck, “A universal scheme for transforming binary algorithms to generate random bits from loaded dice,” *ArXiv:1209.0726 [cs.IT]*, 2012.
- [10] B. Y. Ryabko and E. Matchikina, “Fast and efficient construction of an unbiased random sequence,” *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 1090–1093, May 2000.
- [11] H. Zhou and J. Bruck, “Efficient generation of random bits from finite state Markov chains,” *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 2490–2506, Apr. 2012.
- [12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [13] I. Csiszár, “The method of types,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
- [14] A. Martín, N. Merhav, G. Seroussi, and M. J. Weinberger, “Twice-universal simulation of Markov sources and individual sequences,” *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4245–4255, Sep. 2010.
- [15] R. Shaltiel, “An introduction to randomness extractors,” in *ICALP 2011, Part II*, ser. LNCS, L. Aceto, M. Henzinger, and J. Sgall, Eds., vol. 6756. Springer, Jul. 2011, pp. 21–41.
- [16] T. Linder, V. Tarokh, and K. Zeger, “Existence of optimal prefix codes for infinite source alphabets,” *IEEE Trans. Inform. Theory*, vol. 43, no. 6, pp. 2026–2028, Nov. 1997.
- [17] P. Whittle, “Some distribution and moment formulae for the Markov chain,” *J. Roy. Statist. Soc. Ser. B*, vol. 17, pp. 235–242, 1955.
- [18] L. Goodman, “Exact probabilities and asymptotic relationships for some statistics from m -th order Markov chains,” *Annals of Mathematical Statistics*, vol. 29, pp. 476–490, 1958.
- [19] T. M. Cover, “Enumerative source encoding,” *IEEE Trans. Inform. Theory*, vol. IT-19, no. 1, pp. 73–77, Jan. 1973.

- [20] B. Y. Ryabko, "Fast and efficient coding of information sources," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 96–99, Jan. 1994.
- [21] —, "Fast enumeration of combinatorial objects," *Discr. Math. and its Appl.*, vol. 10, no. 2, pp. 101–110, 1998.
- [22] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 5–20, Jan. 2004.
- [23] N. Merhav, G. Seroussi, and M. J. Weinberger, "Universal delay-limited simulation," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5525–5533, Dec. 2008.
- [24] S.-i. Pae and M. C. Loui, "Randomizing functions: Simulation of a discrete probability distribution using a source of unknown distribution," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 4965–4976, Nov. 2006.
- [25] A. Juels, M. Jakobsson, E. Shriver, and B. K. Hillyer, "How to turn loaded dice into fair coins," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 911–921, May 2000.
- [26] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 384–396, Mar. 1994.
- [27] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.
- [28] D. Baron and Y. Bresler, "An $O(n)$ semi-predictive universal encoder via the BWT," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 928–937, May 2004.
- [29] Á. Martín, G. Seroussi, and M. J. Weinberger, "Linear time universal coding and time reversal of tree sources via FSM closure," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1442–1468, Jul. 2004.
- [30] R. A. Rueppel and J. L. Massey, "Leaf-average node-sum interchanges in rooted trees with applications," in *Communications and Cryptography: Two sides of One Tapestry*, ser. The Springer International Series in Engineering and Computer Science, R. E. Blahut, J. Daniel J. Costello, U. Maurer, and T. Mittelholzer, Eds. Kluwer Academic Publishers, 1994, pp. 343–356.
- [31] T. J. Tjalkens and F. M. Willems, "Variable-to-fixed length codes for Markov sources," *IEEE Transactions on Information Theory*, vol. IT-33, no. 2, pp. 246–257, Mar. 1987.
- [32] —, "A universal variable-to-fixed length source code based on Lawrence's algorithm," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 247–253, Mar. 1992.
- [33] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2104–2109, Sep. 1999.
- [34] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
- [35] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *Annals Math. Stat.*, vol. 42, no. 6, pp. 1992–2007, 1971.
- [36] J. L. Massey, "The entropy of a rooted tree with probabilities," in *Proc. 1983 International Symposium on Information Theory*, St. Jovite, Canada, Sep. 1983.